Insights into the microscale spatial dynamics of dengue and chikungunya in Southeast Asia

by

Henrik Salje

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March, 2014

Abstract

The spatial dynamics of many diseases are generally studied at the macroscale, including the spread of pathogens between countries and continents. Disease dispersal within communities is less well understood. This gap is partly due to a lack of statistical approaches that can accurately characterize spatial and temporal dependence of disease processes in the presence of underlying spatial heterogeneities that can hide any signal. Here we developed approaches that estimate (a) the mean distance between sequential cases in a transmission chain and (b) spatial dependence between cases over different time-frames (irrespective of who infected whom) from point pattern incidence data. Importantly, our approaches are valid where we only observe a tiny fraction of infections and there exist both multiple overlapping transmission chains and spatial heterogeneities in disease surveillance. We demonstrated the robustness of our approaches using simulation. We then applied them to geocoded dengue case data from Bangkok, Thailand, a disease that has been in endemic circulation in this city for decades. We estimated that the mean transmission distance for dengue in the city was 50m (varying between 44m and 64m between 1994 and 2006). Further, the aggregation of short range individual transmissions led to the presence of larger scale spatial temporal dependence, with clustering of all cases within any month observed at distances up to 1km. We also observed patterns of spatiotemporal

ABSTRACT

dependence consistent with the expected impacts of homotypic immunity, heterotypic immunity and immune enhancement of disease at these distances. Our observations indicate that individual transmissions (which encompass both human and mosquito movements) tend to be not be much further than neighboring households, however, immunological memory of dengue serotypes occurs at the neighborhood level in this large urban setting. Infections between neighboring households driving disease spread was also supported for chikungunya, a pathogen transmitted by the same mosquitoes as dengue: we estimated a mean transmission distance of 60m (95% confidence interval: 50m - 70m) from an outbreak of the virus in a village in Bangladesh. The findings presented here have broad implications for understanding the mechanisms of dengue and chikungunya dispersal, the tailoring of intervention measures and the parametrization of mathematical models of disease spread. In addition, the methods presented have wide-ranging application across disease systems.

Thesis Readers

Douglas Norris, PhD, Committee Chair
Professor, Department of Molecular Microbiology and Immunology

Derek Cummings, PhD, Advisor
Associate Professor, Department of Epidemiology

Justin Lessler, PhD
Assistant Professor, Department of Epidemiology

Anna Durbin, MD
Associate Professor, Department of International Health

Acknowledgments

First and foremost I would like to thank Dr. Derek Cummings, my PhD advisor, for providing guidance and support throughout my entire PhD. From the outset he has provided patient and insightful direction to my research and I am extremely fortunate to have worked closely with him. Further, Dr. Justin Lessler has been integral to much of my research and I am grateful for his kind support.

The research presented herein was conducted during my time as part of the Infectious Disease Dynamics group at Johns Hopkins Bloomberg School of Public Health. The group is led by Dr. Cummings and Dr. Lessler and is (in my biased opinion) a world leader in infectious disease dynamics research. It combines methodological development, theoretical and computational approaches and the design and implementation of field studies to help our understanding of disease dynamics and the potential impact of different interventions.

As with most epidemiological research, the work presented here relies on a strong network of collaborators across universities, research institutions, hospitals and ministries of public health. I am indebted to their support. In particular In-Kyu Yoon, Robert Gibbons, Richard Jarman and Ananda Nisalak at the Armed Forces Research Institute of Medical Sciences (AFRIMS) in Bangkok, Thailand have been close collaborators throughout my PhD. AFRIMS has been conducting research in dengue in

ACKNOWLEDGMENTS

Thailand for over 50 years and have been at the center of much of our understanding of the disease. In addition, Dr. Suchitra Nimmannitya at the Queen Sirikit National Institute of Child Health, also in Bangkok, has helped collect dengue patient data for many years. From Bangladesh, I would like to thank Emily Gurley from the International Centre for Diarrhoeal Disease Research, Bangladesh and Mahmudur Rahman from the Institute of Epidemiology, Disease Control and Research. They have been integral to much of the research we have conducted in dengue and chikungunya in Bangladesh. I look forward to continuing to work closely with these institutions in the future.

I would like to thank my fellow doctoral and post-doctoral students in the Infectious Disease Dynamics group. In particular, I would like to thank Isabel Rodriguez-Barraquer, Andrew Azman, Kaitlin Rainwater-Lovett and Ben Althouse for their (often daily) input into my research as well as their friendship. I have no doubt that these individuals will become leaders of infectious disease research in years to come. I would also like to thank Katherine Footer, Nathan Conduit, Joris Nathanson, Nicholas Darcey and Robert Eaglestone for their close friendship over many years, irrespective of differences in physical location. Finally I would like to thank my parents, Lisa and Ekhard and my sisters, Joelle, Jeanne, Lea-Cecile and Barbara for their continued loving support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human disease from mosquito-transmitted viruses remains one of the biggest causes of morbidity and mortality in Southeast Asia, especially among children. By the time they reach adulthood most individuals in the region will have been infected at least twice by dengue, the most common arbovirus in the world [1, 2]. In addition, chikungunya, an arbovirus with the same vectors as dengue, has reemerged in areas where it had not been observed for many years [3]. Unlike malaria, the mosquitoes that transmit dengue and chikungunya bite mainly during the day making bed nets largely redundant and no licensed vaccine currently exists [4]. In addition, the vectors have become well adapted to urban communities. They are usually found inside homes and will oviposit in containers as small as bottle caps, making the targeting of immature forms of the mosquito difficult [2, 5].

Characterizing the spatial dynamics of infectious disease, whether transmitted by mosquitoes or not, is critical to our understanding of the mechanisms of disease spread. Further, it allows us to effectively tailor and implement control measures and to estimate the future course of pathogen dispersal. For example, spatially explicit models allowed the identification of birds as a major factor in the nationwide spread of West Nile virus in the USA and helped inform culling measures in the 2001 foot

and mouth outbreak in the United Kingdom [6, 7].

1.1 Transmission kernels and global spatial dependence

Two related concepts can help us characterize the spatial signature of infectious diseases: transmission kernels and global spatial dependence. Transmission kernels describe the probability distribution function of the distances between sequential cases in a transmission chain. Characterizing transmission kernels is key to understanding how far a newly infected individual is located from the case that infected him or her. Transmission kernels consider the distance between two sequential *human* (or occasionally animal) cases. In vector-borne diseases this single measure will encompass both human and vector movements. Global spatial dependence (often referred to as second-order clustering) captures the tendency for infected individuals to be found near each other, irrespective of who infected whom (Figure 1.1). Global spatial dependence therefore captures distally related cases or unrelated transmission chains circulating in the same area. By contrast *local* spatial clustering (or first order spatial dependence) measures the tendency of cases to occur around a particular point in space. While measures of local clustering have many uses in epidemiology, including the development of risk maps they will not be considered here [8, 9].

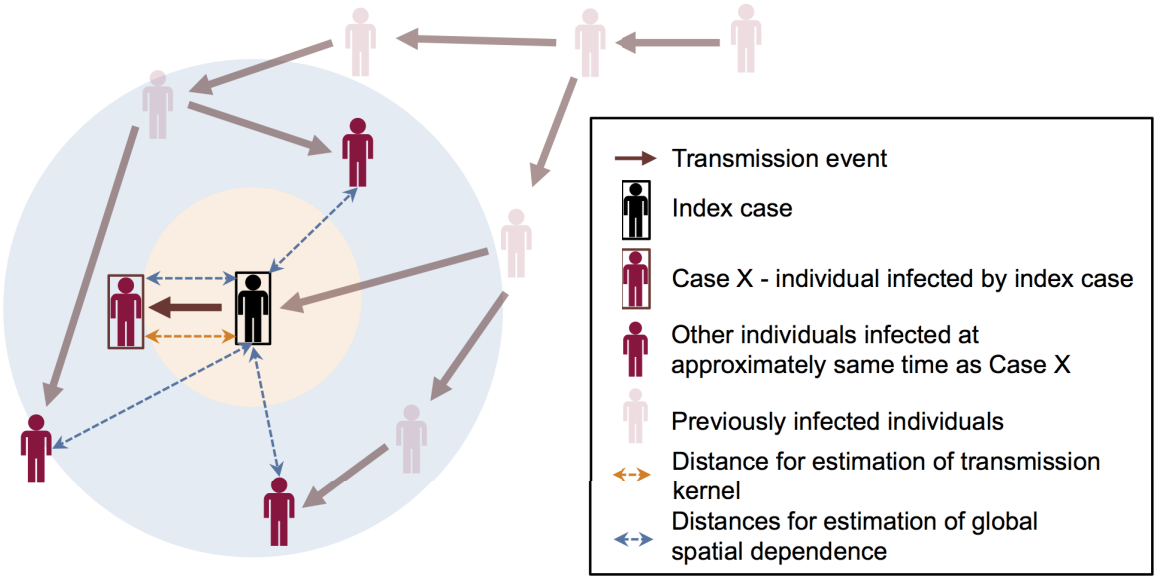


Figure 1.1: Example transmission chain. Each individual represents the location of an infection. The transmission kernel considers only the distance to the immediate subsequent case from an index case whereas global spatial dependence considers all cases and does not consider the relationship between them. Note that we can consider global spatial dependence ignoring time (i.e., all cases irrespective of when they occur) or within specific time windows (such as within a generation time, as illustrated here).

1.2 Existing estimates of the small-scale spatial dynamics of dengue and chikungunya

Previous work has characterized the large-scale disease patterns of dengue, including between continents, countries and across districts within a country [10, 11]. The small-scale dynamics of dengue within communities, however, has not been well described. Cluster investigations using small numbers of index cases in rural communities in Northern Thailand found infections of genetically similar viruses after a delay of 15 days at distances of under 100m [12]. While these findings supported the presence of transmissions over short distances, it is unclear if the observed subsequent

cases were direct transmissions from the index case or more distally related. In addition, the study could not detect infections over large distances. It is also unclear if these findings are consistent with the spatial dynamics of dengue in urban centers, where population movements may be substantially different. A further attempt to characterize the spatial dynamics of dengue was conducted using a cohort in Iquitos, Peru [13]. The study collected location diaries from infected individuals and found that infection risk was correlated by where individuals spent their day (termed 'activity space'). These findings suggested that human movement was a key determinant of the spread of the disease.

As with dengue, the spatial dynamics of chikungunya has been described at several continental and national levels [14, 15]. As far as we are aware, the small-scale global spatial dependence or transmission kernel for chikungunya has not previously been characterized.

1.3 A note on existing statistics

Global spatial dependence

The statistical analysis of spatial point pattern data is an established field, including in relation to the study of infectious diseases, where points generally refer to the coordinates of infected individuals (usually in the form of home location) [16]. Within this field there exist several global spatial dependence statistics that either use the exact locations of cases or aggregate them into grid cells to describe the tendency for infected individuals to be found together. Aggregate (or quadrant) approaches initially place a grid over the study area and count the number of cases falling within

each grid cell. Area-level statistics can then be used to determine if cells with many cases tend to occur near each other (e.g Moran's I or Geary's c) [17,18]. While simple to implement, these methods are highly sensitive to the width of the grid cells and do not provide information of the distribution of cases within any cell. An alternative approach is to calculate the mean distance between each case and its closest neighbor [19]. While the nearest-neighbor approaches do not initially require the grouping of cases and therefore avoid the loss of information from the aggregation of cases, they cannot characterize spatial dependence over different distances. Another distance-based approach, the K-function, by contrast uses estimates of the expected number of cases over a range of distances.

$$K(d) = \lambda^{-1} E[N(d)] \tag{1.1}$$

where λ is the spatial intensity of the cases in the study area (usually calculated as the size of the study area divided by the number of cases) and $E[N(d)]$ is the expected number of cases within distance d of a case. The value of the K-function in a homogenous Poisson process is πd^2 (i.e., where no spatial dependence between cases exists) [16]. Comparisons between the K-function for an observed point pattern and this theoretical value can be used for the detection of spatial dependence. Where there exists spatial heterogeneity in the underlying population at risk, differences between the K-function of cases and the K-function of controls, representative of the underlying population, can be calculated instead.

An important failing of existing measures of spatial dependence, including the K-function, is in their interpretability. The measures are largely constrained to the

detection of presence or absence of spatial dependence within any single setting and cannot be interpreted using classical epidemiological concepts such as relative risk. Measures of spatial dependence can not normally be compared across settings.

Transmission kernels

Transmission kernels can be estimated directly through active follow-up of infected individuals and characterizing who infected whom [6]. However, contact tracing is rarely undertaken: it is resource intensive and even when previously infected contacts are found, it is often difficult to identify direct transmission. Genetic approaches can help confirm whether the infecting pathogens are related in two individuals, however, in rapidly evolving species such as HIV even this can be difficult [20]. The direct detection of infection pairs is further complicated in many diseases by the high proportion of asymptomatic or mildly symptomatic individuals, which may be easily missed. These factors have meant that virtually no transmission kernels have been estimated for infectious diseases. The major exception is diseases in farm animals, where the historic movement for large numbers of livestock are often available [6, 21, 22].

An alternative approach, developed by Keeling et al., is to indirectly measure transmission kernels by finding the kernel that is most consistent with the observed distribution of cases at a single snapshot in time [23]. The approach was able to estimate the transmission kernel for the foot and mouth outbreak in the UK. However, it remains unclear whether it is robust to large numbers of unobserved cases as well as heterogeneities in the observation process or in the underlying population at risk.

1.4 Overview

This dissertation sets out novel methodologies that allow us to measure both global spatial dependence and transmission kernels in realistic scenarios where only passive surveillance data is available and we observe only a tiny fraction of all cases. These approaches are then applied to case data from Thailand and Bangladesh.

The dissertation is divided into two parts. Part one has three manuscripts that describe the spatial dependence and transmission kernel estimates of the homes of hospitalized dengue patients in Bangkok, Thailand and chikungunya cases in Tangail district, Bangladesh. The first manuscript uses two novel methods to characterize spatial dependence when there exists information on the infecting pathogen (in this case serotype). We estimate the short-term spatial dependence between the homes of dengue patients from a children's hospital in Bangkok, a city that has observed endemic dengue transmission for decades. In addition, we explore whether there are localized long-term effects of dengue transmission on future serotype-specific incidence patterns.

The second manuscript explores the relationship between spatial dependence and transmission kernels. The manuscript describes an approach to indirectly estimate the transmission kernel using the mean separation of case pairs. It then estimates both the spatial dependence and the transmission kernel os chikungunya using data from an outbreak investigation from villages in Tangail, Bangladesh in 2013.

The approach described in the second manuscript cannot be used where multiple transmission chains exist and is therefore constrained to outbreak settings. We therefore extend the method in a third manuscript using only pairs of cases that are closest to each other in space over short periods of time and will therefore tend to

come from the same transmission chain. Having demonstrated the robustness of the approach through simulation, we apply it to dengue data from Bangkok between 1994 and 2006.

Part two consists of a single manuscript. We found evidence in chapter 2 that incidence patterns in a location are correlated with future incidence at that location. Immunity patterns within a community could therefore act as a marker of future disease risk at that location. This would require accurate determination of an individual's immune status. The manuscript explores the variability in the most common assay to describe serotype-specific immunity: the Plaque Reduction Neutralization Assay. By using repeated assays on the same serum, we estimated the variance in titers to specific viruses.

References

[1] D. J. Gubler, "Dengue and dengue hemorrhagic fever." *Clinical microbiology reviews*, vol. 11, no. 3, pp. 480–496, Jul. 1998.

[2] S. B. Halstead, "Dengue," Imperial College Press, London, Oct. 2008.

[3] A. M. Powers and C. H. Logue, "Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus." *The Journal of general virology*, vol. 88, no. Pt 9, pp. 2363–2377, Sep. 2007.

[4] M. Yasuno and R. J. Tonn, "A study of biting habits of Aedes aegypti in Bangkok, Thailand." *Bulletin of the World Health Organization*, vol. 43, no. 2, pp. 319–325, 1970.

[5] L. C. Harrington, T. W. Scott, K. Lerdthusnee, R. C. Coleman, A. Costero, G. G. Clark, J. J. Jones, S. Kitthawee, P. Kittayapong, R. Sithiprasasna, and J. D. Edman, "Dispersal of the dengue vector Aedes aegypti within and between rural communities." *The American journal of tropical medicine and hygiene*, vol. 72, no. 2, pp. 209–220, Feb. 2005.

[6] N. M. Ferguson, C. A. Donnelly, and R. M. Anderson, "The foot-and-mouth epi-

demic in Great Britain: pattern of spread and impact of interventions." *Science*, vol. 292, no. 5519, pp. 1155–1160, May 2001.

[7] O. G. Pybus, M. A. Suchard, P. Lemey, F. J. Bernardin, A. Rambaut, F. W. Crawford, R. R. Gray, N. Arinaminpathy, S. L. Stramer, M. P. Busch, and E. L. Delwart, "Unifying the spatial epidemiology and molecular evolution of emerging epidemics." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 37, pp. 15 066–15 071, Sep. 2012.

[8] M. Ali, M. Emch, J. P. Donnay, M. Yunus, and R. B. Sack, "The spatial epidemiology of cholera in an endemic area of Bangladesh." *Social Science & Medicine*, vol. 55, no. 6, pp. 1015–1024, Sep. 2002.

[9] P. Zeman, "Objective assessment of risk maps of tick-borne encephalitis and Lyme borreliosis based on spatial patterns of located cases." *International journal of epidemiology*, vol. 26, no. 5, pp. 1121–1129, Oct. 1997.

[10] O. M. Allicock, P. Lemey, A. J. Tatem, O. G. Pybus, S. N. Bennett, B. A. Mueller, M. A. Suchard, J. E. Foster, A. Rambaut, and C. V. F. Carrington, "Phylogeography and Population Dynamics of Dengue Viruses in the Americas," *Molecular Biology and Evolution*, vol. 29, no. 6, pp. 1533–1543, May 2012.

[11] D. A. T. Cummings, R. A. Irizarry, N. E. Huang, T. P. Endy, A. Nisalak, K. Ungchusak, and D. S. Burke, "Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand," *Nature*, vol. 427, no. 6972, pp. 344–347, Jan. 2004.

[12] M. P. Mammen, C. Pimgate, C. J. M. Koenraadt, A. L. Rothman, J. Aldstadt, A. Nisalak, R. G. Jarman, J. W. Jones, A. Srikiatkhachorn, C. A. Ypil-Butac,

REFERENCES

A. Getis, S. Thammapalo, A. C. Morrison, D. H. Libraty, S. Green, and T. W. Scott, "Spatial and temporal clustering of dengue virus transmission in Thai villages." *PLoS medicine*, vol. 5, no. 11, pp. e205–e205, Nov. 2008.

[13] S. T. Stoddard, B. M. Forshey, A. C. Morrison, V. A. Paz-Soldan, G. M. Vazquez-Prokopec, H. Astete, R. C. Reiner, S. Vilcarromero, J. P. Elder, E. S. Halsey, T. J. Kochel, U. Kitron, and T. W. Scott, "House-to-house human movement drives dengue virus transmission," *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

[14] S. M. Volk, R. Chen, K. A. Tsetsarkin, A. P. Adams, T. I. Garcia, A. A. Sall, F. Nasar, A. J. Schuh, E. C. Holmes, S. Higgs, P. D. Maharaj, A. C. Brault, and S. C. Weaver, "Genome-scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates." *Journal of virology*, vol. 84, no. 13, pp. 6497–6504, Jul. 2010.

[15] D. Taraphdar, A. Sarkar, B. B. Mukhopadhyay, S. Chakrabarti, and S. Chatterjee, "Rapid spread of chikungunya virus following its resurgence during 2006 in West Bengal, India." *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 106, no. 3, pp. 160–166, Mar. 2012.

[16] A. C. Gatrell, T. C. Bailey, P. J. Diggle, and B. S. Rowlingson, "Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology," *Transactions of the Institute of British Geographers, New Series*, vol. 21, no. 1, pp. 256–274, Jan. 1996.

[17] P. Moran, "Notes on continuous stochastic phenomena." *Biometrika*, vol. 37, no. 1-2, pp. 17–23, Jun. 1950.

REFERENCES

[18] R. C. Geary, "The Contiguity Ratio and Statistical Mapping," *The Incorporated Statistician*, vol. 5, no. 3, pp. 115–127+129–146, Nov. 1954.

[19] A. D. Cliff and J. K. Ord, "Model Building and the Analysis of Spatial Pattern in Human Geography," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 37, no. 3, pp. 297–348, Jan. 1975.

[20] S. Resik, P. Lemey, L.-H. Ping, V. Kouri, J. Joanes, J. Pérez, A.-M. Vandamme, and R. Swanstrom, "Limitations to contact tracing and phylogenetic analysis in establishing HIV type 1 transmission networks in Cuba." *AIDS research and human retroviruses*, vol. 23, no. 3, pp. 347–356, Mar. 2007.

[21] C. Nassuato, G. J. Boender, P. L. Eblé, L. Alborali, S. Bellini, and T. J. Hagenaars, "Spatial transmission of Swine Vesicular Disease virus in the 2006-2007 epidemic in Lombardy." *PloS one*, vol. 8, no. 5, pp. e62 878–e62 878, 2013.

[22] J. Turner, R. G. Bowers, and M. Baylis, "Modelling bluetongue virus transmission between farms using animal and vector movements." *Scientific reports*, vol. 2, p. 319, 2012.

[23] M. J. Keeling, S. P. Brooks, and C. A. Gilligan, "Using conservation of pattern to estimate spatial parameters from a single snapshot." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 24, pp. 9155–9160, Jun. 2004.

Part I

Characterizing the spatial dynamics of dengue and chikungunya

Chapter 2

Revealing the microscale spatial signature of dengue transmission and immunity in an urban population

Henrik Salje, Justin Lessler, Timothy P. Endy, Frank C. Curriero, Robert V. Gibbons, Ananda Nisalak, Suchitra Nimmannitya, Siripen Kalayanarooj, Richard G. Jarman, Stephen J. Thomas, Donald S. Burke, and Derek A. T. Cummings

2.1 Abstract

It is well known that the distribution of immunity in a population dictates the future incidence of infectious disease, but this process is generally understood at individual- or macro-scales. For example, herd immunity to multiple pathogens has been observed at national and city levels. However, the effects of population immunity have not previously been demonstrated at scales smaller than the city (e.g., neighborhoods). In particular, to our knowledge no study has demonstrated long-term effects of population immunity at scales consistent with the spatial scale of person-to-person transmission. This gap is partly due to a lack of statistical approaches that can accurately characterize spatial and temporal dependence of disease processes in

the presence of underlying spatial heterogeneities (including reporting biases and the distribution of the underlying population) that can hide any signal. Here, we develop two novel methods that use information on the infecting pathogen (e.g., serotype) to describe both shorter term and longer-term spatio-temporal dependence between cases over and above any dependence due to non-disease processes. We build individual based models to demonstrate the robustness of our statistics to different population structures, seasonal effects and biased reporting. We then apply our approaches to dengue case data from Thailand: we use the location of dengue patients homes in Bangkok over a five-year period with the serotype of the infecting pathogen to investigate the spatiotemporal distribution of disease risk at small spatial scales. We find evidence for localized transmission at distances of under 1km. We also observe patterns of spatiotemporal dependence consistent with the expected impacts of homotypic immunity, heterotypic immunity and immune enhancement of disease at these distances. Our observations indicate that immunological memory of dengue serotypes occurs at the neighborhood level in this large urban setting. These methods have broad application to studying the spatiotemporal structure of disease risk where pathogen serotype or genetic information is known.

2.2 Introduction

Individual risk to infectious diseases is largely determined by immune status and the rate of contact with infectious agents. Past infection or vaccination in an area can reduce infection risk for susceptible individuals by eliminating potentially infectious neighbors [1–6]. Daily movements of host populations determine the spatial scale at which the immunity of neighbors is relevant for disease risk. Models that assume large-scale mixing will have homogenous levels of population immunity [7]; however, there may be important differences at smaller scales (hundreds of meters) that affect the distribution of disease. Analysis of spatiotemporal locations of cases at fine resolutions may reveal the micro-scale dynamics of transmission and population immunity.

Dengue is a viral disease transmitted by the Aedes mosquito with clinical manifestations ranging from asymptomatic illness to potentially fatal dengue haemorrhagic fever [8]. Dengue is present in over 100 countries, causing an estimated 50 million infections and 19,000 deaths each year (WHO 2007). A wide range of vector, human, viral and environmental factors determine the spatial and temporal patterns of dengue infection [8]. These include the spatial distribution and movement of mosquitos and humans; life span, oviposition and blood feeding tendencies of the mosquito; the infectiveness of both hosts; and the spatial distribution of immunity in humans [8]. There are four serotypes of dengue virus (DENV1-4). All four have circulated in Bangkok, Thailand for decades [9]. After infection, individuals develop lifelong immunity to the infecting serotype (homotypic immunity), and there is evidence that they are temporarily protected from infection with other serotypes (heterotypic immunity) [10]. However, once susceptibility to other serotypes returns, these individuals are at increased risk of severe disease upon infection (heterotypic immune enhance-

ment) [11, 12].

Several studies have described the spatial clustering of dengue cases, but did not explore the effect of population immunity [13–15]. To our knowledge, the effect of population immunity on micro-scale disease dynamics has never been systematically characterized using empirical data. This is understandable, as direct observation of the spatial and temporal dynamics of cases and immunity is difficult and resource intensive, requiring longitudinal observation of immune status and case incidence over large spatial and temporal scales. Here, we use a novel approach to characterize the dynamics of population immunity and its effect on future incidence using only the spatiotemporal distribution of clinical dengue cases presenting at a single large hospital.

We use the household location of 1,912 children with laboratory confirmed dengue illness admitted to Queen Sirikit Hospital, Bangkok between 1995 and 2000 to calculate measures of spatiotemporal dependence (Figure 2.1). We use modifications of standard space-time clustering statistics that allow finer resolution of spatiotemporal dependence and control for changes in the underlying spatial and temporal distribution of the population. This approach is built upon an innovative use of the distribution of heterotypic case pairs (those inconsistent with transmission) and homotypic pairs (those consistent with transmission) over a long time-scale to characterize the underlying spatial and temporal heterogeneity in disease risk. We use these methods to investigate whether the spatiotemporal distribution of cases is consistent with localized transmission, the expected effect of long-term homotypic immunity, short-term heterotypic immunity and immune enhancement of disease severity in secondary heterotypic infections.

2.3 Methods

2.3.1 Data collection

Data on clinical cases of dengue between January 1 1995 and December 31 1999 were collected from Queen Sirikit Childrens Hospital in Bangkok, Thailand. There are a total of 2254 cases where address, infecting serotype, and month and year of hospital admission is available. Serotype was determined through reverse transcriptase polymerase chain reaction. Local data managers used base maps for the city to convert addresses to geocoded point locations for each case. One thousand nine hundred and twelve cases were successfully geocoded (85 per cent) (Table 2.1). Original addresses were not available to the analysis team.

	N	N geocoded	% DENV infections
DENV1	571	486	25%
DENV2	474	406	21%
DENV3	1142	964	51%
DENV4	67	56	3%
Total	2254	912	
Secondary infections	1740		77%
DHF	1654		73%
Mean age			7.5 years

Table 2.1: Characteristics of patients with identified serotype from Queen Sirikit Hospital, 1995-1999.

2.3.2 Short-term spatial dependence

To characterize the short-term spatial dependence of homotypic cases within a one month time timeframe we use $\tau(d_1, d_2)$ to calculate the relative probability of a

case occurring during the same month and within distance range d_1 to d_2 of a given case being homotypic, compared to the probability of any other case in that month being homotypic.

$$\tau(d_1, d_2) = \frac{Pr(z_{ij} = 1 | j \in \Omega_i(d_1, d_2))}{Pr(z_{ij} = 1 | j \in \Omega_i(\cdot))} \tag{2.1}$$

where z_i is the serotype of case i and $\Omega_i(d_1, d_2)$ is the set of cases that occur within the same month and within distances d_1 and d_2 of case i.

We estimate τ as:

$$\hat{\tau}(d_1, d_2) = \frac{\sum_{i=1}^{N} \sum_{j \in \Omega_i(d_1, d_2)} z_{ij}}{\sum_{i=1}^{N} |\Omega_i(d_1, d_2)|} \Big/ \frac{\sum_{i=1}^{N} \sum_{j \in \Omega_i(\cdot)} z_{ij}}{\sum_{i=1}^{N} |\Omega_i(\cdot)|} \tag{2.2}$$

where z_{ij} is equal to 1 if the serotype of case i is equal to the serotype of case j and 0 otherwise.

2.3.3 Long-term spatial dependence

To calculate the spatial dependence over several months or years we use $\phi(d_1, d_2, t_1, t_2)$ to calculate the relative probability of a homotypic (or heterotypic) case being within a window of space and time from a case versus that expected if the clustering processes in space and time were independent.

$$\phi_{hom}(d_1, d_2, t_1, t_2) = \frac{Pr(j \in \Omega_i(d_1, d_2, t_1, t_2) | z_i = z_j)}{Pr(j \in \Omega_i(d_1, d_2, \cdot) | z_i = z_j) Pr(j \in \Omega_i(\cdot, t_1, t_2) | z_i = z_j)} \tag{2.3}$$

We estimate $\phi_{hom}(d_1, d_2, t_1, t_2)$ as:

$$\hat{\phi}_{hom}(d_1, d_2, t_1, t_2) = \frac{(\sum_{i=1}^{N} \sum_{j \in \Omega_i(d_1,d_2,t_1,t_2)} z_{ij})(\sum_{i=1}^{N} \sum_{j \in \Omega_i(\cdot,\cdot)} z_{ij})}{(\sum_{i=1}^{N} \sum_{j \in \Omega_i(d_1,d_2,\cdot)} z_{ij})(\sum_{i=1}^{N} \sum_{j \in \Omega_i(\cdot,t_1,t_2)} z_{ij})} \tag{2.4}$$

The function for heterotypic cases is similarly estimated.

Further descriptions of the spatio-temporal dependence methods used and their relationship with existing methodologies can be found in the supporting information.

2.4 Results

2.4.1 Short-term spatial dependence

We find that spatiotemporal dependence exists when the time and location of a case is affected by where and when other cases occur. We characterize the spatial dependence of homotypic cases within a one month time horizon as: the relative probability of a case occurring during the same month and within distance range to of a given case being homotypic, compared to the probability of any other case in that month being homotypic. Both the numerator and denominator are dependent upon the spatiotemporal distribution of cases appearing within the same month regardless of serotype. This formulation thereby controls for underlying heterogeneities in the population that could create spatial or temporal clustering (e.g., variation in population density, hospital and health care utilisation rates, dengue seasonality). Values above one indicate that any two cases that live within the specified distance range of each other were more likely to be homotypic than any two randomly chosen cases presenting during the same month. Cases coming from the same transmission chain

are necessarily homotypic, hence spatial clustering of homotypic cases over short time periods may indicate transmission related cases.

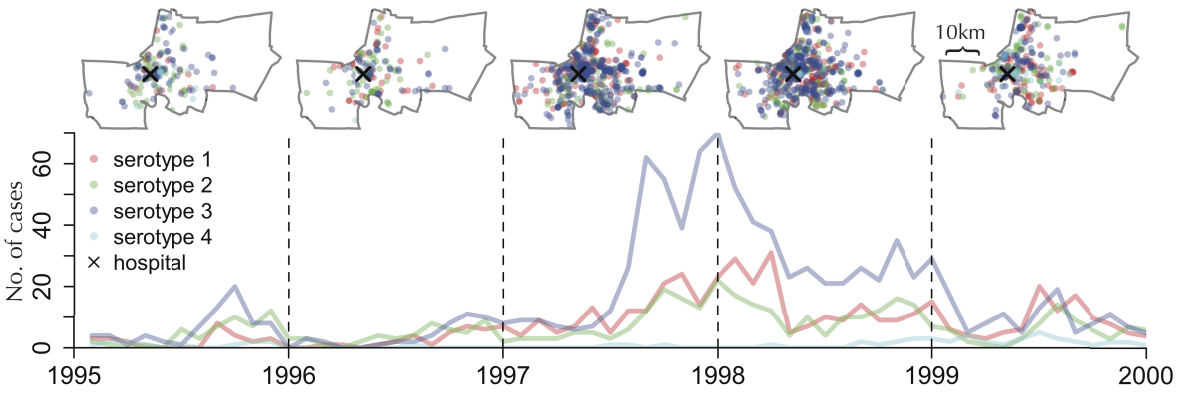


Figure 2.1: Spatial and temporal distribution of clinical cases of dengue disease by month at Queen Sirikit Hospital between 1995 and 2000. The border in each map represents the Bangkok provincial boundary.

We find a 1.82 fold increase in the probability of a case occurring within 200m and the same month of another case being homotypic (95% confidence interval of 1.45, 2.16) (Figure 2.2). This estimate fell to 1.16 (0.97, 1.35) at 1km (250m i.e., the spatial range between 750m and 1.25km). There was an increased probability of cases being homotypic at distances up to 1.8km (250m). However, this was statistically significant only up to 0.7km (250m). Consistent patterns were observed with each of the four serotypes (Figure 2.2). These results suggest that the transmission of dengue in urban Bangkok is focal. Clustering of homotypic cases may be due to local dispersal of host and vector. However, clustering of immune status in the population may contribute to focal case distributions.

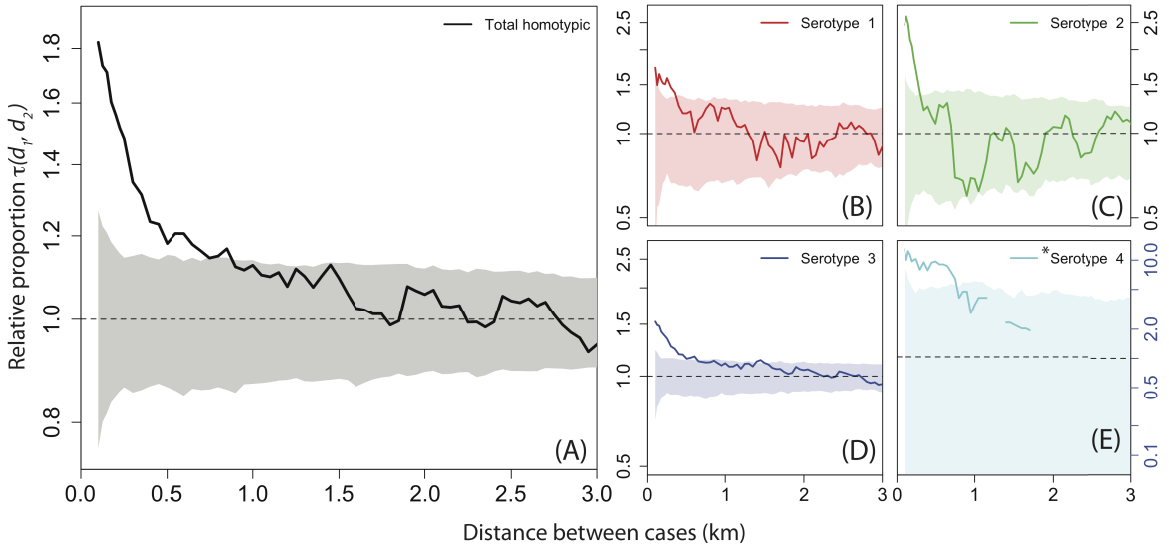


Figure 2.2: Homotypic spatial dependence analysis for cases occurring within the same month. (A) represents the overall homotypic spatial dependence, (B) - (E) represents the homotypic spatial dependence for DENV-1, DENV-2, DENV-3 and DENV-4 cases only, respectfully. The size of the spatial window of analysis $(d_2 - d_1)$ was kept at 0.5km when d_2 was greater than 0.5km. When d_2 was less than 0.5km, d_1 was equal to 0. Estimates are plotted at the midpoint of the spatial range. The shaded area represents 95 per cent intervals of a null distribution generated from 1000 simulations where the time point at which a case occurs is randomly reassigned.

2.4.2 Long-term spatial dependence

The immune profile of the population can induce both short- and long-term spatiotemporal dependence in dengue cases. If we assume neighborhood composition remains mostly the same within the study period and that detected cases are representative of serotype-specific incidence in that neighborhood, we would expect clustering of a particular serotype to result in a reduction in future homotypic cases in that vicinity. Likewise, during the period of short-term cross protection we expect to see fewer heterotypic cases occur near previous dengue cases. Conversely, immune

enhancement may lead to increases in heterotypic cases at longer temporal lags [12].

Spatiotemporally dependent processes are often described using $D_0(d,t)$ which estimates the probability of a point occurring within a spatiotemporal distance of another point, compared to the probability of this occurring due to the independent effects of clustering in space and time [16–19]. $D_0(d,t)$ is a cumulative function, hence can only crudely characterize changing patterns of spatiotemporal dependence. Thus, we derive a related function, $\phi(d_1, d_2, t_1, t_2)$, the relative probability of a homotypic (or heterotypic) case being within a window of space and time from a case versus that expected if the clustering processes in space and time were independent.

Patterns of both homotypic and heterotypic spatiotemporal dependence differ substantially from those seen if we ignored serotype (Figure 2.3). We find that homotypic cases were 1.61 (1.42,1.82) as likely to occur within 400m and 4 months from an incident case than would be expected if the spatial and temporal clustering processes were independent (Figure 2.3B). The relative proportion of homotypic cases fell to 1.14 (1.05, 1.23) at 1km (500m) over the same time frame. This period is followed by a significant reduction in homotypic cases in subsequent months. Homotypic cases were 0.77 (0.67, 0.86) times as likely to occur at temporal lags of 8 to 24 months within 400m, and 0.90 (0.84, 0.96) times as likely at 1km (500m) over the same temporal lags.

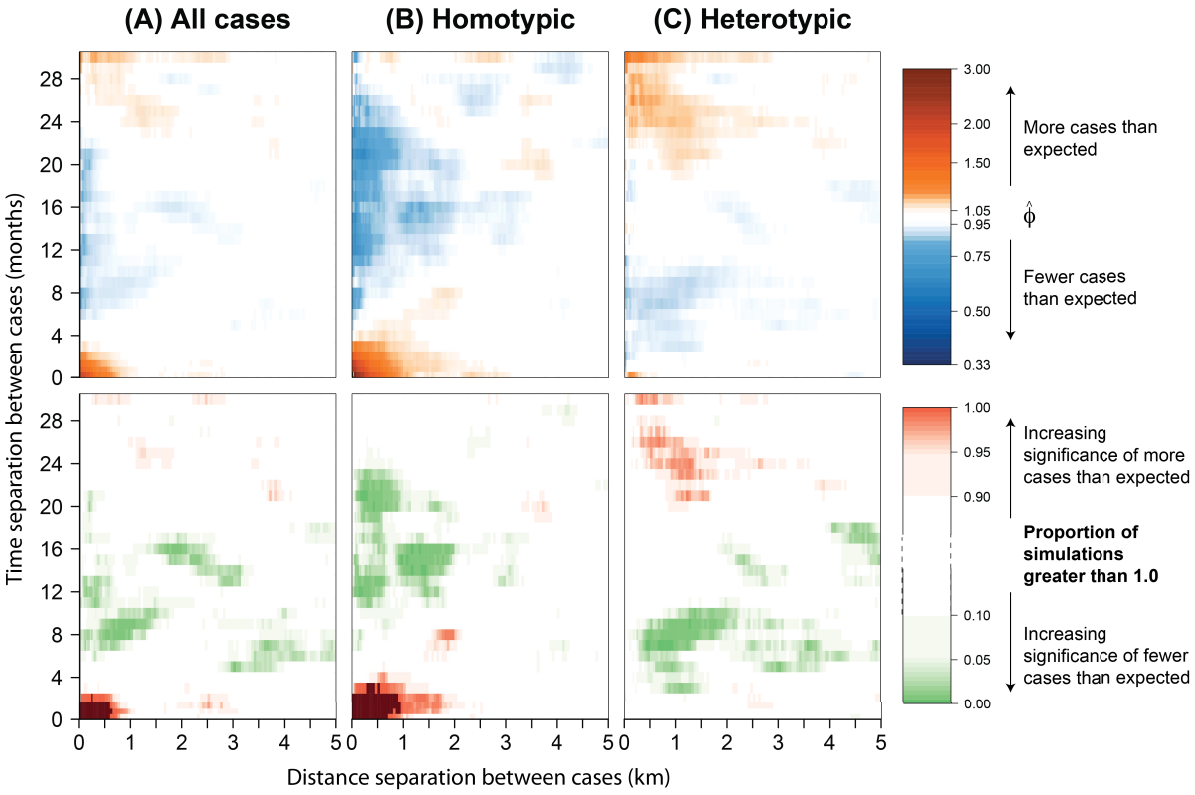


Figure 2.3: Spatial dependence analysis with temporal lags. The average relative proportion of (A) all cases (irrespective of serotype) (B) homotypic and (C) heterotypic cases is illustrated across spatial and temporal lags in the top row. The spatial range $(d_2 - d_1)$ was kept constant at 0.5km when d_2 was greater than 0.5km. When d_2 was less than 0.5km, d_1 was equal to 0. The temporal range $(t_2 - t_1)$ was kept constant at 3 months when t_2 was greater than 3 months. When t_2 was less than 3 months, t_1 was equal to 0. Estimates are plotted at the midpoint of the spatial and temporal ranges. $\phi(d_1, d_2, t_1, t_2)$ estimates under 5% in either direction are colored white. The bottom row sets out the percentage of 1000 bootstrapped simulations that have a $\phi(d_1, d_2, t_1, t_2)$ value greater than 1.0.

We find that heterotypic cases were 0.88 (0.85, 0.96) as likely to occur at 1km (500m) from an incident case at lags of 3 to 10 months (Figure 2.3C). These heterotypic patterns are consistent with Sabin's findings of short-lived cross-protective immunity [10]. Furthermore, there was an increase in heterotypic clustering with a temporal lag of 2 years. Heterotypic cases were 1.11 (1.02, 1.20) times as likely to

occur at temporal lags of 20 to 30 months, 1km (500m) from an incident case. This increase after a period of two years points to elevated risk of disease and supports previous observations of increased disease risk with sequential heterotypic infections [11].

Our analysis includes hospitalized cases only. The spatiotemporal dependence of hospitalized cases is of intrinsic interest. However, the mechanisms that we propose to explain the pattern of dengue cases rely on a correlation between the spatiotemporal distribution of serotypes in hospitalized cases and the spatiotemporal distribution of serotypes in the unobserved cases from which these are drawn. We cannot establish this link with this dataset. However, we use two statistics that give unbiased estimates even in the presence of bias in reporting and heterogeneity in the underlying population density (Appendix A, Figure A.4 and Figure A.5). In addition, we simulated disease transmission processes where there was no spatial dependence between the infected and infecting individuals under conditions of strong seasonal forcing and underlying heterogeneities in the distribution of the population. The phi and tau statistics showed no spatial dependence in these scenarios (Appendix A, Figure A.3). We describe the performance of our statistics in a number of simulated datasets in the supplementary material. We find that our statistics are unbiased under a number of model scenarios. Movement of individuals into or out of the population via birth, death, immigration or emigration may dilute the effect of acquired immunity on future cases. The period of time that a neighborhood remains effectively static will determine, in part, the extent to which we observe spatiotemporal dependence.

2.5 Discussion

These results are from a large georeferenced dataset over several years in an endemic setting where all four serotypes circulate. Such datasets present a rare opportunity to study the dynamics of dengue transmission at a fine spatial resolution. Previous studies that have examined the spatial distribution of dengue have either focused on individual serotypes or have not had fine spatial resolution for large numbers of cases over multiple years [14, 15, 20, 21]. Our study provides further evidence to support the focal nature of DENV, complementing a longitudinal study in Thailand and outbreak investigations in other settings [13, 22, 23]. Our results suggest transmission is spatially local, even in a highly mobile and dense urban population with significant immunity. Our findings that the distribution of cases at one time point, predict the spatial distribution of both homotypic and heterotypic cases at future time points suggests a dispersal mechanism that is partially dictated by the immunity status of the local population. Though the impact of population immunity on pathogen dynamics has been the focus of numerous studies, rarely has the spatiotemporal distribution of cases and importantly, the immunity derived from those cases been shown to predict the future distribution of cases.

Using multiple simulations, we demonstrated that it is unlikely that the observed clustering could be caused by underlying spatial structures, either in the population, or access to the study hospital. In addition, we have shown that our observations could not be generated solely by seasonal dynamics of dengue. Our results provide strong evidence that the clustering process is serotype dependent. We believe that the likeliest and simplest mechanism that would generate serotype specific clustering is the transmission process, which we know to be serotype dependent.

The methods implemented here use variation in pathogen type to characterize the tendency for cases to be found near each other both in the short-term and across temporal lags. We use a passively collected data set to illustrate how, by focusing on differences between event types, such datasets can be used to understand the underlying generating process. These approaches are relevant whenever there exist points of multiple types (e.g., genotype data) or we are interested in changing patterns of spatiotemporal dependence not captured by a cumulative characterisation, regardless of the domain. Here, these methods have revealed micro-scale interactions between transmission, immunity and the future incidence of dengue.

References

[1] R. M. Anderson and R. M. May, *Infectious Diseases of Humans*, ser. Dynamics and Control. Oxford University Press, Aug. 1992.

[2] B. T. Grenfell, O. N. Bjørnstad, and J. Kappey, "Travelling waves and spatial hierarchies in measles epidemics." *Nature*, vol. 414, no. 6865, pp. 716–723, Dec. 2001.

[3] J. Wallinga, P. Teunis, and M. Kretzschmar, "Reconstruction of measles dynamics in a vaccinated population." *Vaccine*, vol. 21, no. 19-20, pp. 2643–2650, Jun. 2003.

[4] N. C. Grassly, C. Fraser, and G. P. Garnett, "Host immunity and synchronized epidemics of syphilis across the United States." *Nature*, vol. 433, no. 7024, pp. 417–421, Jan. 2005.

[5] S. E. Reef, S. B. Redd, E. Abernathy, L. Zimmerman, and J. P. Icenogle, "The epidemiological profile of rubella and congenital rubella syndrome in the United States, 1998-2004: the evidence for absence of endemic transmission." *Clinical Infectious Diseases*, vol. 43 Suppl 3, pp. S126–32, Nov. 2006.

[6] J. P. Fox, L. Elveback, W. Scott, L. Gatewood, and E. Ackerman, "Herd im-

munity: basic concept and relevance to public health immunization practices." *American Journal of Epidemiology*, vol. 94, no. 3, pp. 179–189, Sep. 1971.

[7] R. M. Anderson and R. M. May, "Vaccination and herd immunity to infectious diseases." *Nature*, vol. 318, no. 6044, pp. 323–329, Dec. 1985.

[8] D. J. Gubler, "Dengue and dengue hemorrhagic fever." *Clinical microbiology reviews*, vol. 11, no. 3, pp. 480–496, Jul. 1998.

[9] A. Nisalak, T. P. Endy, S. Nimmannitya, S. Kalayanarooj, U. Thisayakorn, R. M. Scott, D. S. Burke, C. H. Hoke, B. L. Innis, and D. W. Vaughn, "Serotype-specific dengue virus circulation and dengue disease in Bangkok, Thailand from 1973 to 1999." *The American journal of tropical medicine and hygiene*, vol. 68, no. 2, pp. 191–202, Feb. 2003.

[10] A. B. Sabin, "Research on dengue during World War II." *The American journal of tropical medicine and hygiene*, vol. 1, no. 1, pp. 30–50, Jan. 1952.

[11] S. B. Halstead and P. Simasthien, "Observations related to the pathogenesis of dengue hemorrhagic fever. II. Antigenic and biologic properties of dengue viruses and their association with disease response in the host." *The Yale journal of biology and medicine*, vol. 42, no. 5, pp. 276–292, Apr. 1970.

[12] D. S. Burke, A. Nisalak, D. E. Johnson, and R. M. Scott, "A prospective study of dengue infections in Bangkok." *The American journal of tropical medicine and hygiene*, vol. 38, no. 1, pp. 172–180, Jan. 1988.

[13] M. P. Mammen, C. Pimgate, C. J. M. Koenraadt, A. L. Rothman, J. Aldstadt, A. Nisalak, R. G. Jarman, J. W. Jones, A. Srikiatkhachorn, C. A. Ypil-Butac,

REFERENCES

A. Getis, S. Thammapalo, A. C. Morrison, D. H. Libraty, S. Green, and T. W. Scott, "Spatial and temporal clustering of dengue virus transmission in Thai villages." *PLoS medicine*, vol. 5, no. 11, pp. e205–e205, Nov. 2008.

[14] M. A. Rabaa, V. T. Ty Hang, B. Wills, J. Farrar, C. P. Simmons, and E. C. Holmes, "Phylogeography of recently emerged DENV-2 in southern Viet Nam." *PLoS Neglected Tropical Diseases*, vol. 4, no. 7, p. e766, 2010.

[15] J. Raghwani, A. Rambaut, E. C. Holmes, V. T. Hang, T. T. Hien, J. Farrar, B. Wills, N. J. Lennon, B. W. Birren, M. R. Henn, and C. P. Simmons, "Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission." *PLoS pathogens*, vol. 7, no. 6, p. e1002064, Jun. 2011.

[16] A. C. Gatrell, T. C. Bailey, P. J. Diggle, and B. S. Rowlingson, "Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology," *Transactions of the Institute of British Geographers, New Series*, vol. 21, no. 1, pp. 256–274, Jan. 1996.

[17] P. J. Diggle, A. G. Chetwynd, R. Häggkvist, and S. E. Morris, "Second-order analysis of space-time clustering." *Statistical methods in medical research*, vol. 4, no. 2, pp. 124–136, Jun. 1995.

[18] N. P. French, H. E. McCarthy, P. J. Diggle, and C. J. Proudman, "Clustering of equine grass sickness cases in the United Kingdom: a study considering the effect of position-dependent reporting on the space-time K-function." *Epidemiology and infection*, vol. 133, no. 2, pp. 343–348, Apr. 2005.

[19] H. J. Lynch and P. R. Moorcroft, "A spatiotemporal Ripley's K-function to analyze interactions between spruce budworm and fire in British Columbia, Canada," *Canadian Journal of Forest Research*, vol. 38, no. 12, pp. 3112–3119, 2008.

[20] M. J. Schreiber, E. C. Holmes, S. H. Ong, H. S. H. Soh, W. Liu, L. Tanner, P. P. K. Aw, H. C. Tan, L. C. Ng, Y. S. Leo, J. G. H. Low, A. Ong, E. E. Ooi, S. G. Vasudevan, and M. L. Hibberd, "Genomic epidemiology of a dengue virus epidemic in urban Singapore." *Journal of virology*, vol. 83, no. 9, pp. 4163–4173, May 2009.

[21] A. Balmaseda, S. N. Hammond, Y. Tellez, L. Imhoff, Y. Rodriguez, S. I. Saborio, J. C. Mercado, L. Perez, E. Videa, E. Almanza, G. Kuan, M. Reyes, L. Saenz, J. J. Amador, and E. Harris, "High seroprevalence of antibodies against dengue virus in a prospective study of schoolchildren in Managua, Nicaragua," *Tropical medicine & international health : TM & IH*, vol. 11, no. 6, pp. 935–942, Jun. 2006.

[22] S. H. Waterman, R. J. Novak, G. E. Sather, R. E. Bailey, I. Rios, and D. J. Gubler, "Dengue transmission in two Puerto Rican communities in 1982." *The American journal of tropical medicine and hygiene*, vol. 34, no. 3, pp. 625–632, May 1985.

[23] J. B. Siqueira, C. M. T. Martelli, I. J. Maciel, R. M. Oliveira, M. G. Ribeiro, F. P. Amorim, B. C. Moreira, D. D. P. Cardoso, W. V. Souza, and A. L. S. S. Andrade, "Household survey of dengue infection in central Brazil: spatial point pattern analysis and risk factors assessment." *The American journal of tropical medicine and hygiene*, vol. 71, no. 5, pp. 646–651, Nov. 2004.

Chapter 3

Estimating transmission kernels in partially observed epidemics: application to chikungunya in Bangladesh

Henrik Salje, Justin Lessler, Emily Gurley, Mahmudur Rahman and Derek A. T. Cummings

3.1 Abstract

Understanding the typical distance between sequential cases in a transmission chain is critical to elucidating mechanisms of disease dispersal as well as tailoring intervention measures. The distribution of distances between sequential cases, however, has rarely been characterized, as epidemiological investigations that link individual cases are resource intensive or even impossible, especially where an intermediary vector exists. Large numbers of asymptomatic or mildly symptomatic individuals that are not detected by surveillance systems also hinder efforts to detect transmission-linked pairs that would allow the direct estimation of transmission distances. Here we present an approach that uses the spatial and temporal location of cases to indirectly estimate the mean distance between transmission-linked cases. We demonstrate

through simulation that this method recovers the true mean transmission distance even when fewer than five per cent of cases are observed. We apply our approach to an outbreak of chikungunya in Tangail district in Bangladesh where we estimate only 20% of the cases were identified. We estimate that the mean transmission distance between sequential cases is 60 m (95% confidence interval 50 - 70 m). Our approach is applicable across disease systems and can inform intervention and surveillance methods in outbreak settings.

3.2 Introduction

Characterizing the spatial patterns of disease transmission is crucial to understanding the mechanisms of pathogen dispersal as well as to intervention efforts. Despite their usefulness, transmission kernels, the probability distribution of the spatial location of cases (for example home locations as a marker of infection location) in relation to the individuals that infected them, have been difficult to elucidate. We rarely observe infection pairs (i.e., who infected whom) in a transmission network. Where only a minority of cases are observed, analyses tend to be restricted to characterizing the spatial and temporal scales at which cases tend to occur together [1]. The relationship between spatial clustering of cases and a transmission kernel is unclear. Only where we have been able to observe the majority of cases in a transmission network or we have detailed epidemiological data on who infected whom, has estimation of a transmission kernel previously been possible [2]. Here, we present an approach to estimate the transmission kernel using only point locations of cases, times at which individuals become symptomatic and information on the generation time distribution of the pathogen. The method is applicable in situations with full data as well as those

where only a minority of cases is observed. We demonstrate this method using an outbreak of chikungunya in Bangladesh.

Chikungunya is a viral disease transmitted by *Aedes* mosquitos. Around three quarters of infections appear to result in symptoms, which can range from mild fever or rash to debilitating joint pain that can persist for many months [3]. Over the last decade, outbreaks of Chikungunya have been reported throughout much of Southeast Asia [3]. Bangladesh reported its first cases of Chikungunya in 2008 when 32 infected individuals were identified in Chapainawabganj district, and a number of subsequent outbreaks have been reported throughout the country [3, 4]. Despite its widespread presence, the dispersal of the disease is poorly understood. Risk of infection has been associated with farming, especially of rubber trees, however, infections are frequently found among children and individuals not associated with farming [5–7]. Further, the small-scale movement of the virus is unclear. The *Aedes* mosquito does not appear to travel very far, however, human movements could increase the range of infection risk [3, 8]. Elucidating the transmission kernel of the virus could help us understand how the virus moves around communities and in the design of future risk factor studies.

Surveillance for the disease usually relies on case reports in health facilities or on outbreak investigations that systematically look for additional cases once an outbreak has been detected. If the mean distance between transmission pairs was known, we could improve the efficiency of both case finding and surveillance activities. Intervention efforts would also benefit. There is no licensed vaccine currently available so intervention measures (if they exist at all) are reliant on vector control, including insecticide use and the removal of potential ovipositioning sites. Knowledge of the transmission kernel would help the identification of areas at high-risk for exposure

upon the detection of index cases, helping the efficiency of these resource intensive control programs.

Transmission linkage (θ) - The number of transmission events that link two cases (see example in Figure 3.1)

Transmission kernel - The probability distribution function of the distance between sequential cases in a transmission chain

The most recent common ancestor (MRCA) - The most recent infector that can link a pair of cases

Mean transmission distance (μ_k) - The mean of the transmission kernel

Mean distance between θ transmission-linked pairs ($\mu_a(\theta, \mu_k, \sigma_k)$) - The mean distance between cases separated by θ transmission events where the transmission kernel has mean μ_k and standard deviation σ_k

Transmission-linkage weights ($w(\theta, t_1, t_2)$) - The proportion of case pairs where one occurs at t_1 and the other at t_2 that are separated by θ transmission events

Mean distance between all pairs ($\mu_t(t_1, t_2, \mu_k, \sigma_k)$ - The mean distance separating all pairs of cases where one occurs at t_1 and the other at t_2 and the transmission kernel has mean μ_k and standard deviation σ_k

Observed mean distance between case-pairs ($\mu_t^{obs}(t_1, t_2)$ - The observed mean distance separating all pairs of cases where one occurs at t_1 and the other at t_2

Table 3.1: Overview of key terms.

3.3 Methods

If we know where cases occur at a single time point, we can characterize the distribution of distances separating pairs of cases. Where a single transmission chain exists, all case-pairs can be linked through transmission events. For example a case-pair may have been infected by the same infectious individual. Alternatively, their most recent common ancestor (MRCA, to borrow a term from phylogenetic analyses)

may be several generations back (Figure 3.1). Where a constant transmission kernel exists, case-pairs separated by only a few transmission events will tend to be closer together than those separated by many transmission events. The proportion of case-pairs that are separated by a particular number of transmission events will depend on the history of the epidemic: cases early in an outbreak can only be separated by a small number of transmission events whereas the MRCA separating a pair of cases at the end of an epidemic may be large. By comparing the distribution of distances separating case-pairs with the distribution of the number of transmission events separating them we can estimate the mean transmission distance that is most consistent with the observed case distribution.

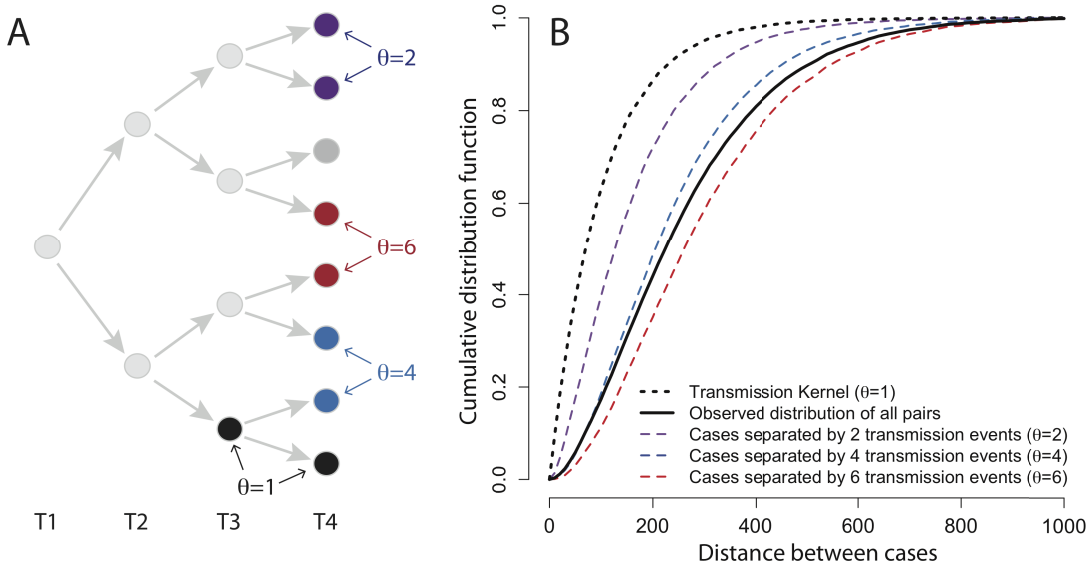


Figure 3.1: (A) Example transmission tree with (B) the cumulative distribution function for pairs of cases separated by different numbers of transmission events assuming a constant exponentially distributed transmission kernel with a mean of 100m.

3.3.1 Mean transmission distance

Where a single transmission chain exists, a pair of cases where one occurs at time point t_1 and the other at time point t_2 can be separated by a number of different possible transmission events (denoted by θ, the number of infection events required to link a pair of cases) (Figure 3.1). For example, two cases occurring at the same time may have been infected by the same infectious individual (in which case $\theta = 2$) or alternatively, the MRCA is further generations back ($\theta = 4$, $\theta = 6$ etc.). If we assume a constant transmission kernel, the directions of transmission events are independent of each other and the distance of transmission events are independent of each other, the distance between pairs of cases will depend on the number of transmission events that separate them. However, without detailed genetic information on the infecting pathogen or contact tracing information, we are unlikely to be able to identify the number of transmission events that separate any two cases. We can, however, calculate the expected distance between all pairs of cases that occur at two time points (the mean of the distribution represented by the solid black line in Figure 3.1B). If we could identify (a) the proportion of case-pairs that are separated by θ transmission events for all possible θ, and (b) the expected distance between pairs of cases separated by θ transmission events under different transmission kernels, we could identify the transmission kernel most consistent with the observed distribution using the following relationship:

$$\mu_t(t_1, t_2, \mu_k, \sigma_k) = \sum_i w(\theta = i, t_1, t_2) \cdot \mu_a(\theta = i, \mu_k, \sigma_k) \tag{3.1}$$

where $\mu_t(t_1, t_2, \mu_k, \sigma_k)$ is the expected distance separating pairs of cases where

one occurs at t_1 and the other at t_2 and the transmission kernel has a mean μ_k and standard deviation σ_k; $\mu_a(\theta, \mu_k, \sigma_k)$ is the mean distance between pairs of cases that are separated by θ transmission events and $w(\theta, t_1, t_2)$ is the proportion of pairs of cases that are separated by θ transmission events (the weights).

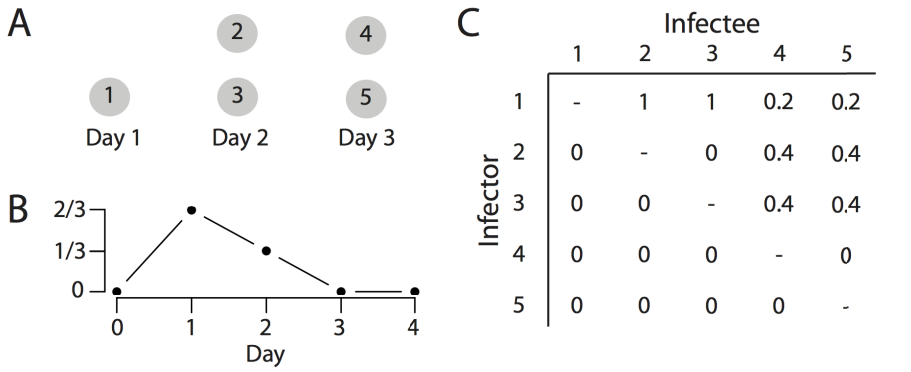
3.3.2 Estimation of weights

To estimate $w(\theta, t_1, t_2)$, we can extend a method developed by Wallinga and Teunis that calculates the probability that a case occurring at time t_1 was infected by a case at time t_2 for all pairs of cases based on the generation time distribution (g, which is assumed known) and the number of cases occurring at each time point [9]. We can produce an n x n matrix, where cell $[i, j]$ represents the probability that a case i was infected by a case infected at the same time point as case j (the Wallinga-Teunis matrix) and n is the total number of cases. For each pair of cases, we can use the Wallinga-Teunis matrix to estimate the probability that they are separated by θ transmission events, by multiplying together the cells of each unique chain (see Figure 3.2 for an example). This assumes that the generation time for all infections are independent of each other. We could compute the probability of every possible path linking two cells, however, this quickly becomes computationally intractable. Instead we can sample transmission trees by randomly choosing the infector for each case, weighted by the probabilities from the Wallinga-Teunis matrix. By re-estimating the tree for each simulation, we adjust for the probability of each transmission tree. Once we have a transmission tree we can compute the number of transmission events required to link each pair of cases. Our estimate of $w(\theta, t_1, t_2)$ is the proportion of simulations in which a case occurring at time t_1 and a case occurring at t_2 are

separated by θ transmission events:

$$\hat{w}(\theta, t_1, t_2) = \frac{\sum_{k=1}^{N_{sim}} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathbf{I_1}(t_i = t_1, t_j = t_2, \Theta_{ij} = \theta)}{N_{sim} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathbf{I_2}(t_i = t_1, t_j = t_2)} \tag{3.2}$$

where N_{sim} is the number of resamples; $\mathbf{I_1}$ is equal to one when case i occurs at time t_1, case j occurs at time t_2 and the simulated transmission tree links case i with case j by θ transmission events (where Θ_{ij} is the number of transmission events linking i and j) and is equal zero otherwise; $\mathbf{I_2}$ is equal to one when case i occurs at time t_1, case j occurs at time t_2 and is equal zero otherwise.



D

Chain	θ	Pr(Chain)
4 - 2 - 1 - 3	3	0.4 x 1 x 1 = 0.4
4 - 1 - 3	2	0.2 x 1 = 0.2
4 - 3	1	0.4
4 - 3 - 1 - 2	3	0.4 x 1 x 1 = 0.4
4 - 1 - 2	2	0.2 x 1 = 0.2
4 - 2	1	0.4
5 - 2 - 1 - 3	3	0.4 x 1 x 1 = 0.4
5 - 1 - 3	2	0.2 x 1 = 0.2
5 - 3	1	0.4
5 - 3 - 1 - 2	3	0.4 x 1 x 1 = 0.4
5 - 1 - 2	2	0.2 x 1 = 0.2
5 - 2	1	0.4

E

θ	wt(θ, t_1= Day 2, t_2= Day 3)
θ = 1	(0.4 + 0.4 + 0.4 + 0.4)/4 = 0.4
θ = 2	(0.2 + 0.2 + 0.2 + 0.2)/4 = 0.2
θ = 3	(0.4 + 0.4 + 0.4 + 0.4)/4 = 0.4

Figure 3.2: Example calculation of the weights from the Wallinga-Teunis matrix. Assume five cases occur over three days as set out in (A) and we know the generation time distribution (B) so that two thirds of sequential infections are a day apart and one third are two days apart. We can build a Wallinga-Teunis matrix (C) that sets out for each case the probability that each other case was its infector. The columns of the matrix have been normalized so that they add to one. (D) Sets out all possible non-zero pathways connecting a case at time 2 with a case at time 3, with the associated number of transmission events (θ) for that chain and the probability of that chain calculated from the Wallinga-Teunis matrix (chains with zero probability such as 4-5-2 have been excluded). (E) sets out the average probability for each θ, which represents the weights used in the calculation of the transmission kernel.

Often only a subset of cases are observed. We can adjust our estimate of $w(\theta, t_1, t_2)$ for partially observed data by randomly sampling with replacement the estimated number of total cases (observed and unobserved) from the observed cases and using

the resampled cases in the calculation of the Wallinga-Teunis matrix. For example if we observed 100 cases but estimate that only 10% of cases are observed, we would sample 1000 cases from the observed cases (with replacement). As we are only interested in the case times for the calculation of $w(\theta, t_1, t_2)$, the locations are unimportant.

3.3.3 Estimation of distance separating cases of known θ

For a transmission kernel with mean μ_k and standard deviation σ_k, we can approximate the mean distance between pairs of cases that are separated by θ transmission events ($\mu_a(\theta, \mu_k, \sigma_k)$) by using the central limit theorem to assume that cases separated by θ transmission events will be approximately normally distributed with a mean (μ_a) and variance (σ_a^2) (Figure 3.1) [10–12].

$$ER^2(\theta, \mu_k, \sigma_k) \approx \mu_k^2 \cdot \theta \left(1 + \frac{\sigma_k^2}{\mu_k^2}\right) \tag{3.3}$$

$$\mu_a(\theta, \mu_k, \sigma_k) \approx 0.5 \cdot \sqrt{\pi \cdot ER^2(\theta, \mu_k, \sigma_k)} \tag{3.4}$$

$$\sigma_a^2(\theta, \mu_k, \sigma_k) = ER^2(\theta, \mu_k, \sigma_k) - \mu_a(\theta, \mu_k, \sigma_k)^2 \tag{3.5}$$

$$\approx ER^2(\theta, \mu_k, \sigma_k)(1 - 0.25\pi) \tag{3.6}$$

where $ER^2(\theta, \mu_k, \sigma_k)$ is the mean squared dispersal distance, μ_k is the mean of the transmission kernel and σ_k is the standard deviation of the transmission kernel.

In situations where the mean and the variance of the kernel are the same, $\mu_a(\theta, \mu_k, \sigma_k)$ and $\sigma_a(\theta, \mu_k, \sigma_k)$ become:

$$\mu_a(\theta, \mu_k, \sigma_k = \mu_k) \approx 0.5 \cdot \mu_k \sqrt{2\pi\theta} \tag{3.7}$$

$$\sigma_a^2(\theta, \mu_k, \sigma_k = \mu_k) \approx 2\theta\mu_k^2(1 - 0.25\pi) \tag{3.8}$$

While this relationship is closest for large θ, we find that for many distributions, this approximation will only cause minor over-estimates of the true mean and variance with small θ (Figure 3.3).

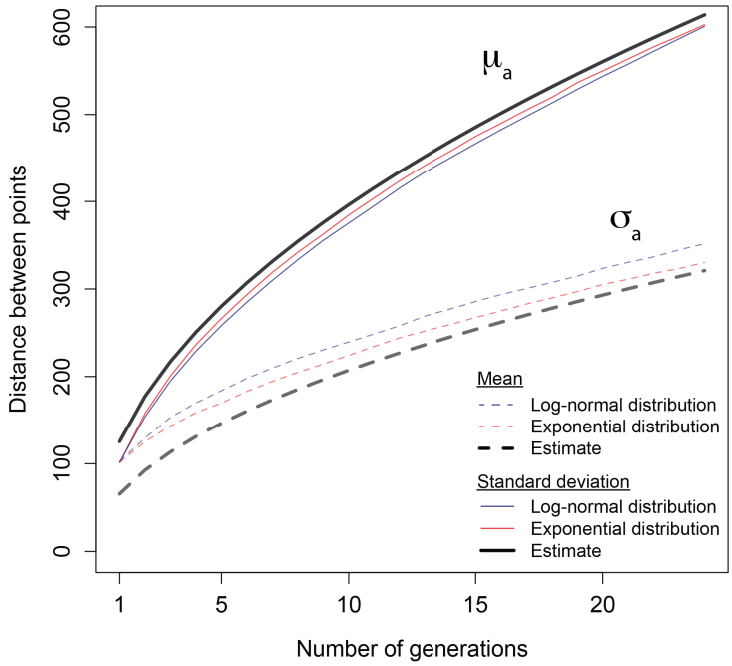


Figure 3.3: Estimates of μ_a (solid lines) and σ_a (dashed lines) from Equations 3.7 and 3.8 compared to simulations with different parametric transmission kernels. All transmission kernels had a mean and standard deviation equal to 100m ($\mu_k = \sigma_k = 100$). The location of a sequential case in a simulated transmission chain was identified by randomly drawing a distance from the paramtric kernel (using either a log-normal or exponential distribution) and a random angle drawn from a uniform distribution between 0 and 2π. The results are the mean and standard deviation of distances between the initial seed (defined as the origin) and the location of the case after n generations (calculated over 20,000 simulations).

3.3.4 Estimation of mean transmission distance where mean and standard deviation of kernel are the same

In situations where the mean and the standard deviation of the transmission kernel are the same, they can be calculated directly for each combinations of t_1 and t_2 as:

$$\hat{\mu_k}(t_1, t_2) = \frac{2\mu_t^{obs}(t_1, t_2)}{\sum_k \hat{w}(\theta = k, t_1, t_2) \cdot \sqrt{2\pi k}} \tag{3.9}$$

where $\mu_t^{obs}(t_1, t_2)$ is the observed mean distance between cases occurring at the two time points. A weighted average estimate across all combinations of t_1 and t_2 is then:

$$\hat{\mu_k} = \hat{\sigma_k} = \frac{1}{\sum_i \sum_j n_{ij}} \sum_i \sum_j \frac{2\mu_t^{obs}(t_1 = i, t_2 = j) n_{ij}}{\sum_k \hat{w}(\theta = k, t_1 = i, t_2 = j) \cdot \sqrt{2\pi k}} \tag{3.10}$$

where n_{ij} is the number of case pairs where one case occurs at time i and one at time j.

3.3.5 Estimation of mean transmission distance where mean and standard deviation of the kernel are different

Where the mean and standard deviation of the transmission differ, we can calculate $\mu_t(t_1, t_2, \mu_k, \sigma_k)$ for different μ_k and σ_k and compare it to $\mu_t^{obs}(t_1, t_2)$.

$$SSE(\mu_k, \sigma_k) = \sum_i \sum_j \left(\mu_t(t_1 = i, t_2 = j, \mu_k, \sigma_k) - \mu_t^{obs}(t_1 = i, t_2 = j)\right)^2 \cdot n_{ij} \tag{3.11}$$

where $SSE(\mu_k, \sigma_k)$ represents the weighted sum of squared errors. The μ_k and σ_k with the corresponding lowest sum of squared errors represents the best estimate of the mean and standard deviation of the transmission kernel.

3.3.6 Confidence intervals

We can generate confidence intervals by bootstrapping the point locations: for example we can re-estimate the mean transmission distance over 1000 resamples and calculate the 2.5 and 97.5 percentile of the resulting distribution.

3.3.7 Performance using simulated data

To assess the performance of our approach we simulated transmission chains with known mean transmission distance. Each chain was generated by initially placing a point (the first case) at the origin at time 0. This case generated daughter infections at a time randomly drawn from a log-normal distribution with a mean of 14 days and standard deviation of 2 days, reflecting the generation time distribution. The location of each daughter case was determined by randomly drawing a direction of infection using a uniform distribution between 0 and 2π and a distance using a transmission kernel with a mean and standard deviation of 100m. The time and the location of each case were recorded. Each daughter case then became a new infector and generated further cases. Each simulation was run for 10 generations.

We ran different scenarios varying (a) the functional form of the transmission kernel (either an exponential distribution or a normal distribution) and (b) the presence of seasonal forcing (either the number of daughter cases per individual was fixed at two or varied seasonally so that the number of cases was drawn from a Poisson distribution with mean $1 + 0.5sin(\pi/2 + 2\pi t/365)$ where t is the time in days).

We assessed the ability of our scenarios to correctly identify the true mean transmission distance under conditions of partially observed data: for each simulation, we

randomly deleted between 0% and 99% of cases before estimating the transmission distance (500 simulations in all). In addition we explored the sensitivity of our results to large misspecification of the mean generation time: we estimated the mean transmission distance where we assumed a mean time of 7 days between sequential infections and where we assumed a mean time of 28 days between sequential infections. Finally, we explored the impact of not adjusting the Wallinga-Teunis matrix for the proportion of unobserved cases (i.e. assuming that all cases were observed when they were not).

3.3.8 Outbreak of Chikungunya in Tangail district, Bangladesh

We applied our technique to an outbreak of chikungunya in Bangladesh. In August 2012, an outbreak of chikungunya in the villages of Palpara, Uttar Gopalpur and North Golpalpur in Tangail district was reported to the Institute of Epidemiology, Disease Control and Research, the Bangladesh governmental center of disease control for outbreak investigation and response. The three villages are connected with no spatial separation between them. In collaboration with the icddr,b, an outbreak investigation team was sent to the outbreak villages. The team went to every house and identified suspected chikungunya cases. A suspected case consisted of fever with either rash or joint pain. The estimated date of fever onset for suspected cases, ages, occupation and all home locations were recorded. All suspected cases that consented also had blood taken, which was tested for evidence of recent chikungunya infection using IgM ELISA. Those that tested positive were considered confirmed cases. A sample of individuals with no symptoms was also tested for evidence of recent chikungunya infection. Informed consent was obtained from all participating

individuals. In all, the outbreak team interviewed 1,970 individuals. They found that 447 had suffered recent symptoms consistent with chikungunya (Table 1). IgM ELISA testing was performed on 246 individuals with symptoms and 172 without. The probability of being a confirmed case given you had symptoms was 70% and the probability of being negative given you had no symptoms was 69%. Applying these proportions to the total number of individuals with and without symptoms gave a total estimated total number of symptomatic cases in the population of 313 and the total number of asymptomatic cases of 472, indicating an attack rate of 40% and an asymptomatic proportion of 60%. The confirmed symptomatic cases represented 22% of all cases. Females were more likely to have symptoms consistent with chikungunya than males (25% vs 20%). The proportion of people infected was consistent across age groups (Appendix B, Figure B.1).

Characterization of clustering of cases

We characterized the spatial dependence observed between all confirmed cases by estimating $\tau(d_1, d_2)$: the probability of an individual becoming infected if he or she lived between distances d_1 and d_2 of another case that occurred within a month of another case that got infected within the previous month relative to the probability of any two individuals becoming infected at that time [1]. All locations of individuals were based on the location of their primary residence. Confidence intervals were generated by 500 bootstrap resamples.

Estimation of mean transmission distance

We estimated the mean transmission distance for chikungunya in this outbreak using the confirmed cases only. As chikungunya is a vector-transmitted disease, the distance represents the mean separation between sequential human cases and will be made up of both human and mosquito movements. We used a lognormal generation time distribution with a mean of 14 days. This estimate is derived from the time from inoculation to peak viremia in *Aedes albopictus* mosquitoes (6 days), the mean time from inoculation to symptom onset in humans (3 days) and the time from symptom onset to peak viremia in humans (5 days) [3, 13–16]. We assumed that the standard deviation of the generation time distribution was 3 days and that the mean and standard deviation of the transmission kernel were the same. To explore the sensitivity of our results to misspecification of the mean generation time, we also estimated the mean transmission distance using a mean generation time of 7 days and a mean generation time of 28 days. In addition, we estimated the mean transmission distance using all suspected cases.

3.4 Results

3.4.1 Performance of approach using simulated data

To assess our ability to estimate the mean transmission distance, we simulated epidemics with a known transmission kernel (exponential distribution with mean and standard deviation of 100m) (Figure 3.4A and Figure 3.4B). We then identified the mean and standard deviation of the transmission kernel that minimized the sum of squared errors (Equation 3.11). We found that a range of values were equally

consistent with the simulated epidemics, including the correct one (red dot in Figure 3.4C). Central limit theorem implies that while the distribution function can look different for cases separated by one generation, they may converge quickly to the same distribution after a handful of generations. For example the following kernels were equally as likely: (1) exponential distribution with mean 100m (red point in Figure 3.4C and red lines in Figures 3.4D), (2) uniform distribution with mean 110m (green), (3) normal distribution with mean 20m and standard deviation 120m (purple) and (4) normal distribution with mean 120m and standard deviation 20m (orange). Whereas a normal distribution with mean 200m and standard deviation 500m (grey) was not supported by the SSE. While the cumulative distribution functions of the four equally likely transmission kernels look slightly different (Figure 3D top), the distributions of pairs of distances separated by only five generations are virtually indistinguishable from each other. However, they are substantially different to the kernel that was not supported by the SSE. Therefore the spatial distribution of cases in epidemics will look similar across spatial kernels as long as the mean and the standard deviation of the kernel lie within the same band in Figure 3.4C.

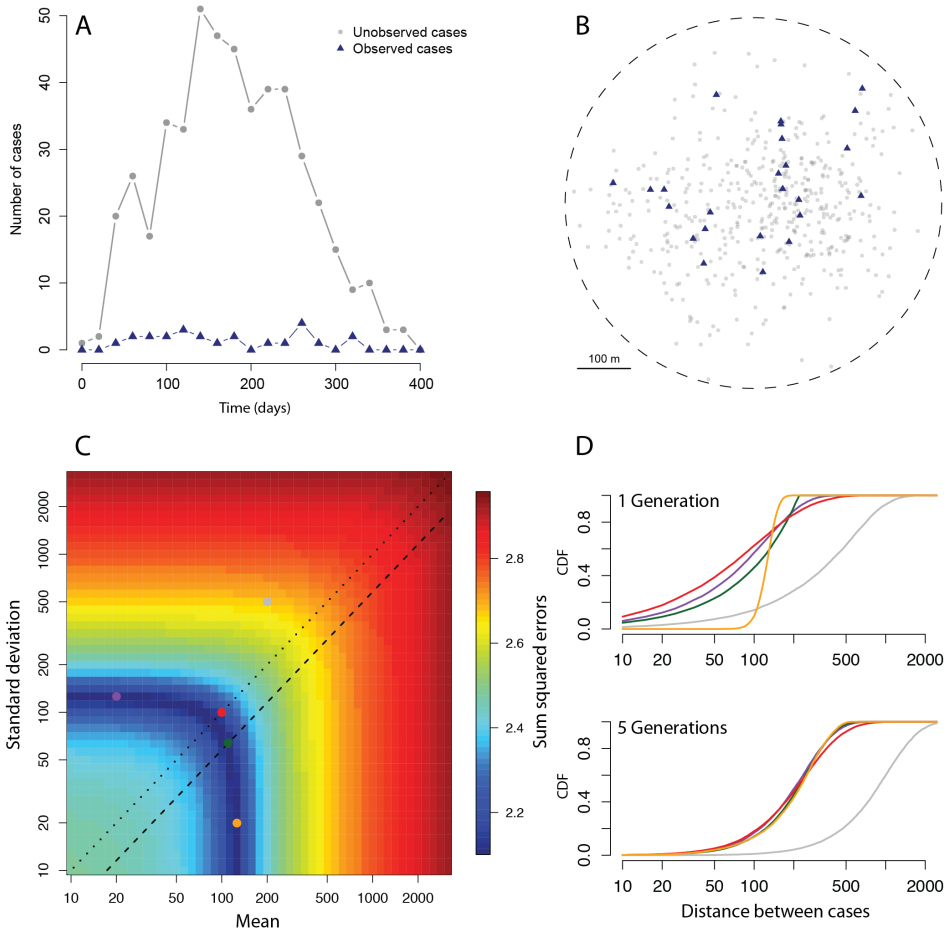


Figure 3.4: Estimates of the mean transmission kernel using simulated data. A transmission chain with an exponentially distributed transmission kernel was simulated and then randomly sampled so that only five per cent of the cases were observed. (A) The epidemic curve for the unobserved and observed cases and their spatial locations (B). (C) The sum of squared errors for different mean and standard deviation of the kernel. The red dot represents the true transmission kernel ($\mu_k = 100m$ and $\sigma_k = 100m$). The dotted line represents kernels where the mean and the standard deviation of the kernel are the same (e.g., exponential distribution). The dashed line represents transmission kernels with a uniform distribution. (D) Illustration of how kernels with similar transmission kernels (top panel) converge to appear identical after only a few generations (bottom panel). The colors of the lines correlate to the points in panel (C).

We were able to capture the true mean transmission distance where the transmission kernel was exponentially distributed, irrespective of the proportion of cases

observed (mean of 98m, 95% confidence interval of 75 - 130) (Figure 3.5). The results were virtually identical for a Gaussian transmission kernel (101m, 81 - 129). In addition, seasonally forced models were also able to identify the true mean transmission distance, even when only a small proportion of cases were observed (101m, 81 - 129). Misspecification of the true mean generation time resulted in small errors in the mean transmission distance estimates: a 100% overestimate of the time between infections resulted in a mean estimate of 128 (99 - 173) whereas a 50% underestimate resulted in a mean distance estimate of 82 (64 - 107). Failing to adjust for the proportion of unobserved cases resulted in an over-estimate of the mean transmission distance when a small proportion of cases were observed. Where fewer than 20% of cases were observed, failing to adjust for the proportion of cases observed resulted in mean estimate of 134m (98 - 194) (Figure 3.5, red curve).

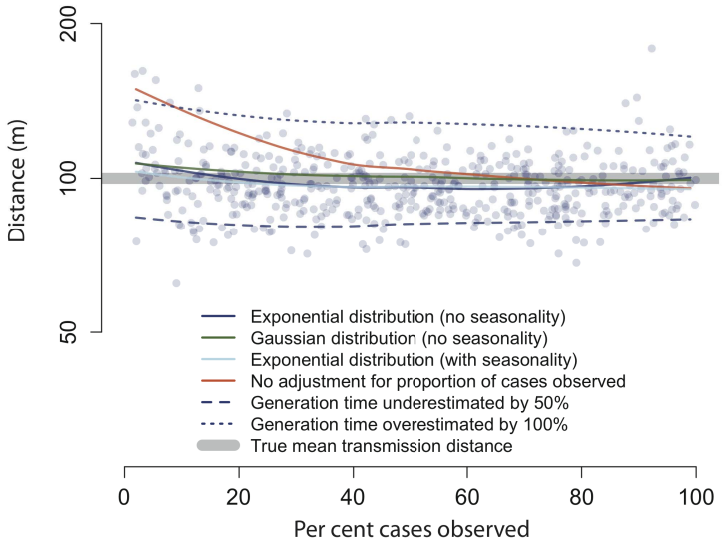


Figure 3.5: Estimates of mean transmission distance from simulated transmission chains where only a subset of cases are observed. Unless otherwise stated, all simulations adjusted the Wallinga-Teunis matrix for the proportion of cases observed. The blue dots represent estimates from individual simulations using a single transmission chain with an exponential distributed transmission kernel and no seasonality. The lines represent loess curves from 500 simulations.

3.4.2 Transmission kernel of chikungunya in Tangail district, Bangladesh

Having demonstrated the robustness of our approach with simulated data, we estimated the transmission kernel for chikungunya in an outbreak in Tangail district, Bangladesh in 2012. Only confirmed cases were used in the estimation of the transmission kernel.

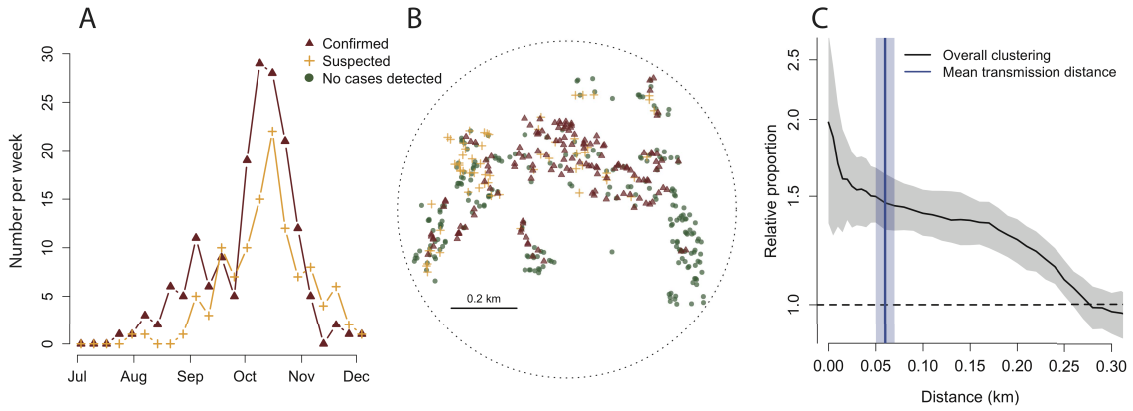


Figure 3.6: (A) Epidemic curve of chikungunya outbreak in Bangladesh. (B) Home locations of confirmed cases, suspected cases (with no confirmed cases) and where no cases were found. (C) Overall clustering of cases using $\tau(d_1, d_2)$ and the estimate of the mean transmission distance with 95% confidence intervals.

The epidemic curve and location of case homes is set out in Figure 3.6. We characterized the spatial dependence between cases occurring within a month of each other. We found that individuals in the study area were 1.5 times more likely to be a case if they lived within 50m of another case occurring within a month relative to the probability of any two individuals being a case at that time (95% confidence interval of 1.4 - 1.7). Spatial dependence between cases occurring within a month of each other was observed at distances up to 300m. We estimated that the mean transmission distance was only 60m (95% confidence intervals of 50m - 70m), demonstrating that small scale transmission distances can give rise to larger scale spatial dependence of cases due to the aggregate correlation of multiple transmission events. If infections were randomly distributed throughout the outbreak villages, the mean transmission distance would have been 380m, reflecting the mean distance between all homes. Our results were robust to substantial differences in the mean generation time: using a mean time of 7 days between sequential infections reduced the mean distance estimate

to 43m (38m - 50m) whereas increasing it to 28 days gave an estimate of 74m (67m - 85m). Using all suspected cases also gave a consistent estimate of 60m (54m - 67m).

3.5 Discussion

We have presented a novel approach to estimate the mean distance separating a case of an infectious disease and the cases that she or he infects in settings where only a minority of cases is observed. We've demonstrated the utility of this technique for both directly and vector-borne pathogens. Through simulation, we demonstrated the robustness of our approach where fewer than five per cent of cases in an outbreak were observed. We then applied it to an outbreak of chikungunya in Bangladesh, an arbovirus that presents a substantial public health burden throughout much of the tropics. We found that a mean distance of not much farther than neighboring households between sequential cases was most consistent with the observed distribution of cases. This has important consequences for spatially targeted interventions including the focused spraying of insecticides.

Two types of *Aedes* mosquitoes, *Aedes albopictus* and *Aedes aegypti* are known to transmit the virus. It has been suggested that *Aedes albopictus* was responsible for the major chikungunya outbreaks in neighboring West Bengal, India, however, the situation is less clear in Bangladesh, where both mosquito types are abundant [17]. *Aedes aegypti* has a limited flight range with mark, release and recapture experiments finding that the majority of mosquitoes tend to stay either in the home from which they were released or immediately next-door [8]. *Aedes albopictus* may have a slighter larger flight range [18]. Human movements around the biting times of the mosquito may also play a key role in the dispersal of the virus. It has previously been ob-

served that while people may travel large distances any given day, movements are centered around their home [19]. This may especially be true during mornings and evenings, important biting times of the mosquito [20]. Our estimate represents the mean distance between the homes of sequentially infected individuals. Our findings also indicate that the home itself is likely to be a good marker of infection location. If infections were commonly occurring elsewhere, such as a school, market or local fields, it would require the unlikely scenario that neighboring households regularly visited the same locations but households farther apart than 60m did not. The role of homes as a marker for infection location is further supported by the substantial number of infected women, who tend to stay in and around the home during the day. We also found no clear differences in the probability of reporting symptoms by age.

While the mean distance between infection pairs was little farther than the separation between homes, significant spatial dependence between cases infected within a month of each other was observed at distances up to 300m. Global spatial dependence captures pairs of cases that are separated by several transmission events as well as direct transmissions and therefore represents the wider risk of cases occurring at the around the same time rather than the distance of transmission-linked cases. Both measures are useful. In immediate responses to outbreaks we are usually interested in identifying areas at high risk of exposure upon detection of index cases rather than individuals directly at risk from the index case. In these situations, the overall clustering of cases will provide a better indication of where infections may be happening, especially in poorly observed epidemics. For example, if mosquito control efforts could be deployed in future outbreaks in similar settings, our findings indicate they should focus on areas up to 300m from the households of detected cases. The mean transmission distance, by contrast, can provide insight on mechanisms that can

be driving the epidemic, useful for longer-term control efforts and can also inform mathematic models of how diseases may spread in similar settings.

We are unable to differentiate between different functional forms of the transmission kernel, however, understanding the mean transmission distance provides a useful indicator of disease spread. We have also demonstrated that different forms of the transmission kernel converge after just a few generations. We also estimated a single transmission distance for chikungunya, assuming that the standard deviation of the transmission distance was equal to the mean. Where the transmission distances are more variable, our estimate will represent an over-estimate of the true mean transmission distance, indicating an even narrower transmission kernel for chikungunya in this setting. Chikungunya has a relatively short generation time. It is unclear how this approach would perform with diseases with much longer or highly variable generation times. Finally, our method requires that all cases in an outbreak are transmission related (even if the MRCA is several generations back). Where more than one transmission chain exists and we have no ability to differentiate between the different chains, such as in endemic settings, we would not be able to use this approach.

In conclusion, we present an approach to estimate the mean transmission distance that will be applicable across disease systems where only a minority of cases is observed. Chikungunya dispersal appears to be driven by small scale house to house transmission.

References

[1] H. Salje, J. Lessler, T. P. Endy, F. C. Curriero, R. V. Gibbons, A. Nisalak, S. Nimmannitya, S. Kalayanarooj, R. G. Jarman, S. J. Thomas, D. S. Burke, and D. A. T. Cummings, "Revealing the microscale spatial signature of dengue transmission and immunity in an urban population." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 24, pp. 9535–9538, Jun. 2012.

[2] N. M. Ferguson, C. A. Donnelly, and R. M. Anderson, "The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions." *Science*, vol. 292, no. 5519, pp. 1155–1160, May 2001.

[3] J. E. Staples, R. F. Breiman, and A. M. Powers, "Chikungunya fever: an epidemiological review of a re-emerging infectious disease." *Clinical Infectious Diseases*, vol. 49, no. 6, pp. 942–948, Sep. 2009.

[4] F. I. Chowdhury, A. Kabir, A. Das, and S. M. Mukerrama, "Chikungunya Fever: An Emerging Threat to Bangladesh," *Journal of Medicine*, vol. 13, pp. 60–64, 2012.

[5] A. F. Yusoff, A. N. Mustafa, H. M. Husaain, W. M. Hamzah, A. M. Yusof, R. Harun, and F. N. Abdullah, "The assessment of risk factors for the Cen-

tral/East African Genotype of chikungunya virus infections in the state of Kelantan: a case control study in Malaysia," *BMC Infectious Diseases*, vol. 13, no. 1, p. 211, 2013.

[6] S. Wangchuk, P. Chinnawirotpisan, T. Dorji, T. Tobgay, T. Dorji, I.-K. Yoon, and S. Fernandez, "Chikungunya fever outbreak, Bhutan, 2012." *Emerging Infectious Diseases*, vol. 19, no. 10, pp. 1681–1684, Oct. 2013.

[7] R. Ansumana, K. H. Jacobsen, T. A. Leski, A. L. Covington, U. Bangura, M. H. Hodges, B. Lin, A. S. Bockarie, J. M. Lamin, M. J. Bockarie, and D. A. Stenger, "Reemergence of chikungunya virus in Bo, Sierra Leone." *Emerging Infectious Diseases*, vol. 19, no. 7, pp. 1108–1110, Jul. 2013.

[8] L. C. Harrington, T. W. Scott, K. Lerdthusnee, R. C. Coleman, A. Costero, G. G. Clark, J. J. Jones, S. Kitthawee, P. Kittayapong, R. Sithiprasasna, and J. D. Edman, "Dispersal of the dengue vector Aedes aegypti within and between rural communities." *The American journal of tropical medicine and hygiene*, vol. 72, no. 2, pp. 209–220, Feb. 2005.

[9] J. Wallinga and P. Teunis, "Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures." *American Journal of Epidemiology*, vol. 160, no. 6, pp. 509–516, Sep. 2004.

[10] P. M. Kareiva and N. Shigesada, "Analyzing Insect Movement as a Correlated Random Walk," *Oecologia*, vol. 56, no. 2/3, pp. 234–238, Jan. 1983.

[11] E. A. Codling, M. J. Plank, and S. Benhamou, "Random walk models in biology." *Journal of the Royal Society, Interface / the Royal Society*, vol. 5, no. 25, pp. 813–834, Aug. 2008.

[12] P. Bovet and S. Benhamou, "Spatial analysis of animals' movements using a correlated random walk model," *Journal of theoretical biology*, vol. 131, no. 4, pp. 419–433, Apr. 1988.

[13] M. Dubrulle, L. Mousson, S. Moutailler, M. Vazeille, and A.-B. Failloux, "Chikungunya virus and Aedes mosquitoes: saliva is infectious as soon as two days after oral infection." *PloS one*, vol. 4, no. 6, pp. e5895–e5895, 2009.

[14] N. Wauquier, P. Becquart, D. Nkoghe, C. Padilla, A. Ndjoyi-Mbiguino, and E. M. Leroy, "The acute phase of Chikungunya virus infection in humans is associated with strong innate immunity and T CD8 cell activation." *The Journal of infectious diseases*, vol. 204, no. 1, pp. 115–123, Jul. 2011.

[15] R. S. Lanciotti, O. L. Kosoy, J. J. Laven, A. J. Panella, J. O. Velez, A. J. Lambert, and G. L. Campbell, "Chikungunya virus in US travelers returning from India, 2006." *Emerging Infectious Diseases*, vol. 13, no. 5, pp. 764–767, Apr. 2007.

[16] K. Rudolph, J. T. Lessler, R. M. Moloney, B. Kmush, and D. Cummings, "Incubation periods of mosquito-borne viral infections: a systematic review (in press)," *The American journal of tropical medicine and hygiene*, Feb. 2014.

[17] D. Taraphdar, A. Sarkar, B. B. Mukhopadhyay, S. Chakrabarti, and S. Chatterjee, "Rapid spread of chikungunya virus following its resurgence during 2006 in West Bengal, India." *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 106, no. 3, pp. 160–166, Mar. 2012.

[18] F. F. Marini, B. B. Caputo, M. M. Pombi, G. G. Tarsitani, and A. A. della Torre, "Study of Aedes albopictus dispersal in Rome, Italy, using sticky traps

in mark-release-recapture experiments." *Medical and Veterinary Entomology*, vol. 24, no. 4, pp. 361–368, Dec. 2010.

[19] G. M. Vazquez-Prokopec, D. Bisanzio, S. T. Stoddard, V. Paz-Soldan, A. C. Morrison, J. P. Elder, J. Ramirez-Paredes, E. S. Halsey, T. J. Kochel, T. W. Scott, and U. Kitron, "Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment." *PloS one*, vol. 8, no. 4, p. e58802, 2013.

[20] M. Yasuno and R. J. Tonn, "A study of biting habits of Aedes aegypti in Bangkok, Thailand." *Bulletin of the World Health Organization*, vol. 43, no. 2, pp. 319–325, 1970.

Chapter 4

Dengue in Bangkok: estimating transmission distances in the presence of multiple overlapping transmission chains

Henrik Salje, Justin Lessler, In-Kyu Yoon, Robert Gibbons, Richard Jarman and Derek A. T. Cummings

4.1 Abstract

Dengue is the most widespread arbovirus in the world. Without a licensed vaccine, intervention efforts rely on resource intensive measures, including the spraying of homes upon detection of an index case. Effective deployment of such spatially targeted interventions requires understanding of the typical distances between sequential human cases in a transmission chain. Such information is critical to understanding the mechanisms that can cause the observed spread of the disease. Elucidating mean transmission distances has been hampered by the presence of numerous overlapping transmission chains, frequent in endemic settings, with no ability to discriminate between chains. Further, the presence of many asymptomatic cases and poor surveillance capabilities result in only a tiny minority of all cases being observed. Here we

present a novel method that is able to accurately estimate the true mean transmission distance using nearest neighbor distances. We demonstrate the robustness of our approach using simulation. Even when fewer than five percent of cases are observed and there exists many overlapping transmission chains, we were able to capture the true mean transmission distance. We estimated the transmission distance of dengue using 8,620 geocoded cases of dengue from Bangkok between 1994 and 2006. We found a remarkably consistent mean transmission distance of 50m (range of 44m to 54m within any one year) over the study period. Our results were robust to broad model misspecification. These findings indicate that small scale movements, not much farther than neighboring households, are driving the dispersal of the disease. These findings will help the tailoring of intervention methods and help inform mathematical models of disease spread. The presented approach is applicable to multiple disease systems.

4.2 Introduction

Understanding the distance between sequential cases in a transmission chain is crucial for infectious disease epidemiology. Both the elucidation of transmission mechanisms and the tailoring of spatially targeted interventions require knowledge of the spatial scales at which transmissions are occurring. In the previous chapter, we presented a method that was able to accurately identify the mean transmission distance in outbreak settings even when only a tiny minority of all cases was observed. By using the mean distance between pairs of cases presenting at two time points and information on the generation time distribution of the disease, we estimated that the mean transmission distance of chikungunya in an outbreak in Bangladesh was

60m. This finding supported a significant role for small-scale transmission between neighboring homes in chikungunya outbreaks. Unfortunately, we cannot apply this approach where multiple transmission chains exist without being able to identify individual chains. For example, dengue virus has circulated for decades with many overlapping transmission chains circulating in any one location, with no ability to differentiate between the chains [1].

Dengue is a arbovirus that has been endemic in Southeast Asia for decades [2,3]. All four serotypes of the virus (DENV1, DENV2, DENV3 and DENV4) can cause severe disease or even death [1, 4]. However, the majority of infections tend to be asymptomatic or result in only minor symptoms, so even the most sophisticated passive surveillance systems will only detect a minority of cases. No licensed vaccine currently exists and the recent disappointing results from the most advanced candidate suggest that the rollout of an effective vaccine may still be many years away [5]. Intervention measures therefore center around vector control, including the spraying of insecticides or removal of potential ovipositioning sites [1]. Such measures are resource intensive and tend to be performed upon detection of cases through passive surveillance. Vector control teams in Bangkok are often deployed to spray around the homes of cases that present at one of the city hospitals. There is little evidence that these interventions are effective. One challenge that may limit there effectiveness is the accurate targeting of control to areas where transmission ongoing. Understanding the typical distance between sequential cases in a transmission chain is therefore critical for effective use of these interventions. However, understanding who infected whom is particularly difficult in vector transmitted diseases as the virus may pass between unrelated individuals, hampering the use of traditional contact tracing approaches.

Here we present a method that uses only the distance to the spatially closest observed case from any index case, which we assume will tend to be from the same transmission chain. We explore the robustness of our approach using simulated data and then apply it to dengue case data from Bangkok.

4.3 Methods

4.3.1 Distribution of distances between cases at two time points

As set out in the previous chapter, the distances between pairs of cases occurring at two time points will depend on the number of transmission events (θ) that separate them. If we know the proportion of case-pairs at two time points that are separated by each possible θ, we can estimate the overall mean distance of all case pairs as a weighted sum.

$$\mu_t(t_1, t_2, \mu_k, \sigma_k) = \sum_i w(\theta = i, t_1, t_2) \cdot \mu_a(\theta = i, \mu_k, \sigma_k) \tag{4.1}$$

where $\mu_t(t_1, t_2, \mu_k, \sigma_k$ is the mean distance separating all pairs of cases where one occurs at t_1 and the other at t_2; $\mu_a(\theta, \mu_k, \sigma_k)$ is the mean and $\sigma_a(\theta, \mu_k, \sigma_k)$ is the standard deviation of the distance between cases separated by θ transmission events where the transmission kernel has mean μ_k and standard deviation σ_k and $w(\theta, t_1, t_2)$ is the proportion of case pairs occurring at t_1 and t_2, respectively that are separated by θ transmission events.

The variance of the distance between all case pairs (σ_t^2) can be similarly estimated.

$$\sigma_t^2(t_1, t_2, \mu_k, \sigma_k) = \sum_i w(\theta = i, t_1, t_2) \cdot \big((\mu_a(\theta = i, \mu_k, \sigma_k) - \mu_t(t_1, t_2, \mu_k, \sigma_k))^2 + \sigma_a^2(\theta = i, \mu_k, \sigma_k)\big) \quad (4.2)$$

When the mean and the standard deviation of the transmission kernel are the same, the estimates become:

$$\mu_t(t_1, t_2, \mu_k, \sigma_k = \mu_k) \approx \sum_i w(\theta = i, t_1, t_2) \cdot 0.5 \cdot \mu_k \sqrt{2\pi i} \quad (4.3)$$

$$\approx \mu_k \sum_i w(\theta = i, t_1, t_2) \cdot 0.5 \cdot \sqrt{2\pi i} \quad (4.4)$$

$$\approx \mu_k C_1 \quad (4.5)$$

where:

$$C_1 = \sum_i w(\theta = i, t_1, t_2) \cdot 0.5 \cdot \sqrt{2\pi i} \quad (4.6)$$

$$\begin{aligned}\sigma_t^2(t_1,t_2,\mu_k,\sigma_k) &= \sum_i w(\theta=i,t_1,t_2)\cdot\Bigg((0.5\cdot\mu_k\sqrt{2\pi i}-0.5\sum_j w(\theta=j,t_1,t_2)\mu_k\sqrt{2\pi j})^2 \\ &\qquad + 2i\mu_k^2(1-0.25\pi)\Bigg) \\ &= \mu_k^2\sum_i w(\theta=i,t_1,t_2)\cdot\Bigg((0.5\cdot\sqrt{2\pi i}-0.5\sum_j w(\theta=j,t_1,t_2)\sqrt{2\pi j})^2 \\ &\qquad + 2i(1-0.25\pi)\Bigg) \\ &= \mu_k^2 C_2 \qquad (4.7)\end{aligned}$$

where:

$$\begin{aligned}C_2 = \sum_i w(\theta=i,t_1,t_2)\cdot\Bigg((0.5\cdot\sqrt{2\pi i}-0.5\sum_j w(\theta=j,t_1,t_2)\sqrt{2\pi j})^2 \\ + 2i(1-0.25\pi)\Bigg)\end{aligned}$$

We previously demonstrated how we can use the Wallinga-Teunis matrix to estimate $w(\theta,t_1,t_2)$ (Equation 3.2) (Figure 3.2).

As pairs of cases separated by a particular θ are approximately Gaussian distributed, all pairwise distances between cases at two time points can be approximated as a mixture of Gaussian distributions [6]. However, as distances cannot be negative, instead we can use a Weibull distribution (a related distribution which can only take positive values) with mean $\mu_t(t_1,t_2,\mu_k,\sigma_k)$ and variance $\sigma_t^2(t_1,t_2,\mu_k,\sigma_k)$. The performance of the Weibull distribution in describing the distribution of all distances is set out in Figure 4.1.

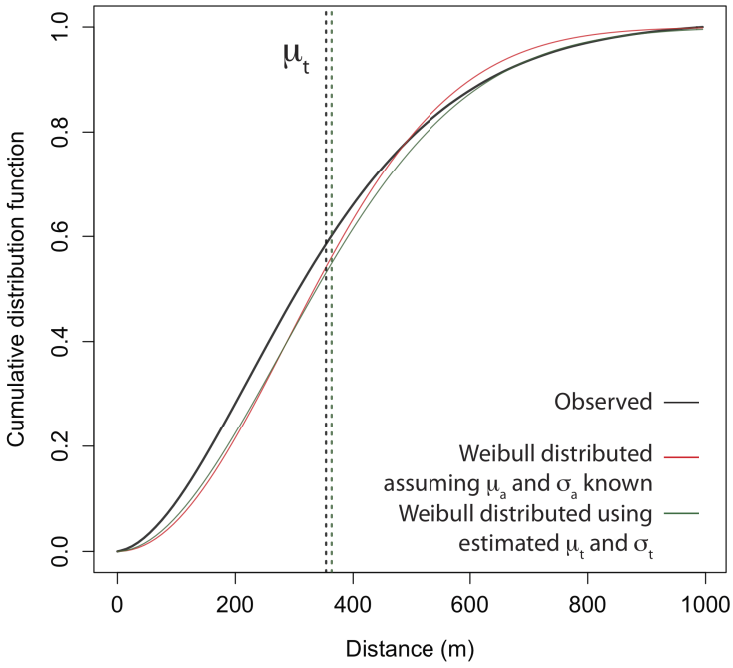


Figure 4.1: Distribution of distances between all pairs of cases after six generations from transmission chain simulations using a exponentially distributed transmission kernel with mean of 100m. Each chain was started with one case at the origin. Each case generated further cases, with the number of daughter cases determined using a Poisson distribution with a mean of two (i.e. a constant effective reproductive number of two). The black line is the cumulative distribution function of all distances occurring in the sixth generation ($t_1 = t_2 = 6$). The red line was calculated using a Weibull distribution where the true mean (μ_t) and standard deviation (σ_t) were assumed known. The green line was calculated using a Weibull distribution where μ_t and σ_t were estimated using the proportion of cases separated by each θ with the weights for each θ estimated using the Wallinga-Teunis matrix.

4.3.2 Minimum expected distance between cases at two time points

If there exists a number of circulating transmission chains at any time point, the distribution of distances between all pairs of points will include case-pairs of unrelated transmission chains, resulting in misleading (and usually over-estimates) of the true transmission distance. If we instead assume that only the *closest* case to any index case will tend to be from the same transmission chain, we can use the expected minimum distance of the Weibull distribution described above. This will require we approximately know the average proportion of cases that belong to the same chain (η).

We are therefore interested in the expected distance between a case occurring at t_1 and the spatially closest case at t_2 (defined here as $M(t_1, t_2, \mu_k, n)$, where n is the number of observed cases at t_2 that come from the same chain). A benefit of approximating the distribution of distances via a Weibull distribution is that the expected minimum of a Weibull distribution of n observations can be simply derived [7].

$$M(t_1, t_2, \mu_k, n) = \gamma n \Gamma\left(1 + 1/k\right) n^{-(1+1/k)} \tag{4.8}$$

where $\Gamma(x)$ represents the gamma function and k is calculated by iteratively solving:

$$\left(\frac{\sigma_t}{\mu_t}\right)^2 = \left(\frac{\Gamma(1+2/k)}{\Gamma(1+1/k)}\right)^2 - 1 \tag{4.9}$$

and γ is calculated as:

$$\gamma = \frac{\mu_t}{\Gamma(1 + 1/k)} \tag{4.10}$$

We can estimate n as the number of cases occurring at t_2 multiplied by the proportion of observed cases that come from the same chain (η). We assessed the performance of the Weibull function in estimating the expected minimum distance using simulated data (Figure 4.2).

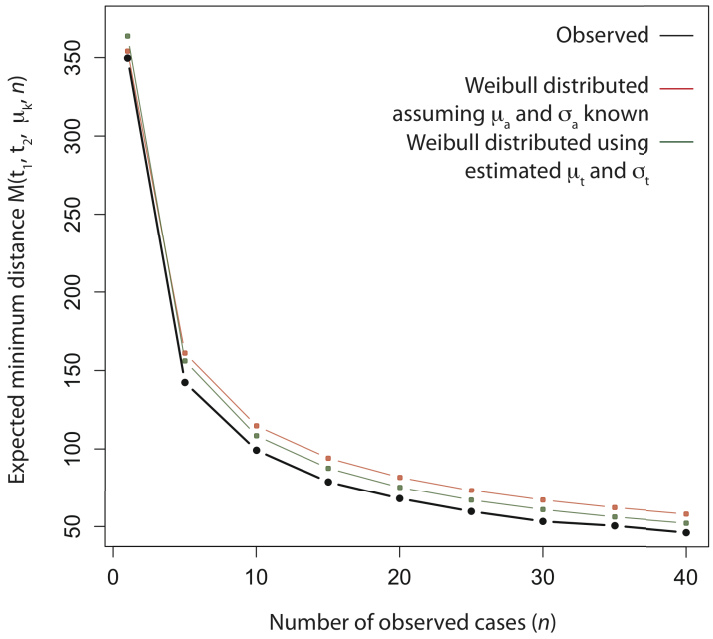


Figure 4.2: The observed (from simulation, black line) and shortest distance of case-pairs when only a subset (n) of cases are observed and that expected from a Weibull distribution (as calculated using Equation 4.8) when either μ_t and σ_t are assumed known (red line) or when they are estimated (green line).

4.3.3 Estimating the mean transmission distance

We now have a framework where we can estimate the expected minimum distance between pairs of cases under different mean transmission distances (μ_k) where one occurs at t_1 and the other at t_2 where we also know (1) the generation time distribution of the pathogen (used in the calculation of the Wallinga-Teunis matrix), (2) the approximate proportion of all cases that are observed (used in the calculation of the Wallinga-Teunis matrix) and (3) the proportion of case-pairs that come from the same transmission chain (η). We can compare the observed mean shortest case-pair distance ($M_{ob}(t_1, t_2)$, where t_1 and t_2 are specified to some level of precision, e.g. days) with that expected under different μ_k for all t_1, t_2 combinations and identify the one most consistent with the data using the weighted sum of squared differences.

$$SSE(\mu_k) = \frac{\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} n_j \left(M(t_1 = i, t_2 = j, \mu_k, n = n_j\eta) - M_{ob}(t_1 = i, t_2 = j)\right)^2}{\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} n_j} \tag{4.11}$$

where n_j is the number of cases at t_2 when $i \neq j$ and the number of cases at t_2 minus one when $i = j$ (to avoid self comparisons).

4.3.4 Use of truncated distances to help estimation of η

It is difficult to estimate the proportion of case-pairs that are from the same chain (or its reciprocal, the number of circulating transmission chains) over a wide area, such as over hospital catchment areas. We can instead consider the proportion of case-pairs within a maximum distance d_{max} of any case. In this approach, n_j is the

expected number of cases at t_2 within d_{max} of a case at t_1. Note that the value of d_{max} should remain significantly larger than potential values of the mean transmission distance. The use of truncated areas has the added benefit of reducing the impact of heterogeneous observation. Even if there exists significant variability across a study region in the probability of a case being observed, if the probability of being observed within d_{max} is constant, the estimate of μ_k will be unchanged.

4.3.5 Assessing performance using simulated data

To assess the performance of our approach we simulated transmission chains with a known kernel and then estimated the mean transmission distance for each simulation. The transmission chains were generated by randomly introducing cases (seeds) onto a surface. Each case then generated daughter cases, representing transmission events. The time between the initial seeds and daughter cases was drawn from a log-normal distribution with a mean of 14 days and standard deviation of three days representing the estimated generation time of dengue (the time separating subsequent human cases in a transmission chain, including time in vector) [8]. The number of daughter cases for each individual was either fixed at two or seasonally forced. The distance between infector and daughter cases was determined by an exponential transmission kernel with mean and standard deviation both equal to 100m. The angle of the transmission event was drawn from a uniform distribution between 0 and 2π. The process was then repeated with the daughter cases generating new offspring. Altogether there were 10 generations conducted. We simulated either a single transmission chain or 20 different chains. Where several chains were simulated, all chains were seeded at the same time. To simulate different levels of overlap between the chains, we varied the area of the

surface on which the initial seeds were randomly placed so that (on average) either 10%, 25% or 50% of cases within 100m of any case were from the same transmission chain.

We explored the impact of a range of factors on our ability to recover the true mean transmission distance: the functional form of the transmission kernel (exponential or normally distributed), seasonal variability in the effective reproductive number, the number of transmission chains and the extent of overlap between transmission chains (Table 4.1). We also explore the impact of misspecification of the proportion of cases that originate from the same chain (η). Finally we estimated the impact of a spatially heterogeneous observation process to recreate the impact of cases farther away from a hospital being less likely to be observed. We performed 500 simulations for each scenario.

	Kernel distribution	Seasonal forcing (1)	No. chains	Chain overlap (2)	Observation process (3)	η estimate
(a)	Exponential	No	1	-	Homogeneous	Accurate
(b)	Exponential	No	20	50%	Homogeneous	Accurate
(c)	Exponential	No	20	25%	Homogeneous	Accurate
(d)	Exponential	No	20	10%	Homogeneous	Accurate
(e)	Normal	No	20	10%	Homogeneous	Accurate
(f)	Exponential	Yes	20	10%	Homogeneous	Accurate
(g)	Exponential	No	20	10%	Homogeneous	Overestimate by 100%
(h)	Exponential	No	20	10%	Homogeneous	Underestimate by 50%
(i)	Exponential	No	20	10%	Heterogeneous	Accurate

Table 4.1: Overview of different scenarios of simulated transmission chains. All simulated transmission kernels had a mean and standard deviation of 100m. (1) Effective reproductive number when seasonal forcing applied calculated as $R_t = 1 + 0.5sin(\pi/2 + \pi t/365)$. When no seasonal forcing, the reproductive number was fixed at 2. (2) The proportion of case pairs within 100m that are from different chains. (3) When the observation process was homogeneous, between 1% and 100% of cases were randomly chosen to be included in the analysis. When the observation process was heterogeneous, the probability of a case being observed was determined by $e^{-5d_i/D}$, where d_i was the distance from the origin to case i and D is the distance to the farthest case. This represents reduced probability of being observed the farther away from the surveillance site (e.g., hospital).

4.3.6 Estimation of the mean transmission distance of dengue in Bangkok

Data collection

We estimated the mean transmission distance of dengue using cases of confirmed dengue that presented at Queen Sirikit National Institute of Child Health (QSNICH), a large children's hospital in the center of Bangkok, Thailand between 1994 and 2006. Approximately one in ten of all hospitalized dengue cases in Bangkok present at this hospital [9]. For each case, the home address of the patient was recorded and subsequently geocoded using detailed base maps of the city. Where possible the infecting serotype was also identified using RT-PCR.

Estimation of number of circulating transmission chains

To use our method to estimate the mean transmission distance of dengue in this setting we need an estimate of the average proportion of cases that come from the same transmission chain (η). We can use the serotype distribution of cases to approximate η. Heterotypic cases (cases caused by different serotypes) cannot come from the same transmission chain (e.g. a DENV1 will always have been infected by another DENV1 case). Further, if we consider that homotypic cases (cases caused by the same serotype) more than 5km apart are also caused by different transmission chains we can estimate the underlying probability of a pair of transmission unrelated cases being caused by the same serotype (η).

$$P_0 = \frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathbf{I}(t_{ij} < T, d_{ij} > 5km, z_{ij} = 1)}{\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathbf{I}(t_{ij} < T, d_{ij} > 5km)} \tag{4.12}$$

where $\mathbf{I}$ is an indicator variable, t_{ij} is the absolute time difference between cases i and j; d_{ij} is the distance between them and z_{ij} is equal to one when the serotypes of i and j are the same and 0 otherwise. For the purposes of this analysis, T was taken to be equal to one month.

The expected probability of a pair of cases being of the same serotype at a distance x (P_x) is then:

$$P_x = \eta + P_0(1 - \eta) \tag{4.13}$$

Solving for η gives us:

$$\eta = \frac{P_x - P_0}{1 - P_0} \tag{4.14}$$

The number of transmission chains within x is the reciprocal of η. P_x can be measured in our data as the proportion of pairs of cases at distance x that are of the same serotype. x was taken to be 1km, a distance larger than the expected mean transmission distance of dengue.

Estimation of the mean transmission distance

We calculated the mean transmission distance for dengue for each year between 1994 and 2006 using all cases (whether a serotype was available or not) that lived within 10km of the hospital. We assumed that the probability of a case being observed was 1% (ie. $\rho = 0.01$). We used a log-normal generation time distribution with a mean of 14 days and standard deviation of three days [8]. To maximize the probability that the closest case-pairs at two points was from the same transmission chain, only case-pairs occurring within a month of each other were used (i.e. the maximum difference between t_1 and t_2 was one month). Thus new transmission chains moving into an area several months after a previous chain had circulated in the same area would not bias our results.

Sensitivity analyses

To assess the sensitivity of our results to our input conditions we conducted sensitivity analyses varying the mean generation time, maximum distance from the hospital, proportion of cases observed, maximum analysis distance (d_{max}) and the proportion of cases that come from the same chain within d_{max}. For each input parameter in turn, we changed the baseline parameter value to the minimum value and then the maximum value in the sensitivity range set out in Table 4.2.

Parameter	Baseline (sensitivity range)	Source
Mean generation time	14 days (10 - 20)	[8, 10, 11]
Maximum distance from hospital	10km (5 - 100)	Model assumption
Proportion observed (ρ)	0.01 (0.005 - 0.03)	Model assumption
Maximum analysis distance (d_{max})	1km (0.5 - 2)	Model input
Proportion of case-pairs from same chain within d_{max} (η)	0.38 (0.25 - 0.50)	Estimated from serotype data

Table 4.2: Key parameter values and sensitivity ranges.

4.4 Results

4.4.1 Simulated data

To explore the robustness of our approach we simulated transmission chains with a true mean transmission distance of 100m and randomly thinned the cases. We found that where there exists only a single transmission chain, our approach resulted in a mean estimate of 101m (95% confidence interval of 56 - 132) with minimal difference by proportion of cases observed (Figure 4.3). In situations of 20 overlapping transmission chains where 25% of cases within 100m from any index case were part of unrelated transmission chains, the mean transmission distance was estimated at 94m (82 - 108). Where greater overlap between chains exists and 50% of cases came from

different chains, the approach slightly underestimated the true transmission distance with a mean estimate of 68m (56 - 76). If the proportion of case-pairs coming from the same chain was overestimated by 100%, the mean transmission distance was slightly higher than the true value (126m, 114 - 140) whereas if it was underestimated by 50%, it was lower than the true value (mean transmission distance of 70m, 60 - 77). There were no differences in the functional form of the transmission kernel, with the results for the Gaussian kernel the same as the exponential kernel. Finally in situations of spatially heterogeneous observation, where there was no truncation of the analysis distance (i.e. $d_{max} = \infty$), the mean transmission distance was slightly underestimated (mean distance of 79m, 95% confidence interval of 57 - 93), however, limiting the analysis to cases within 1km only resulted in unbiased estimates (mean distance of 99m, 95% confidence interval of 80 - 128). Our approach therefore appears able to estimate the mean transmission distance where there is no ability to differentiate between different chains, even in settings with substantial overlap between chains and where only a minority of cases are observed. Further, even where our estimates was biased, the magnitude of the bias was small and the estimate of the mean transmission distance would still be a useful indicator of the typical transmission distance.

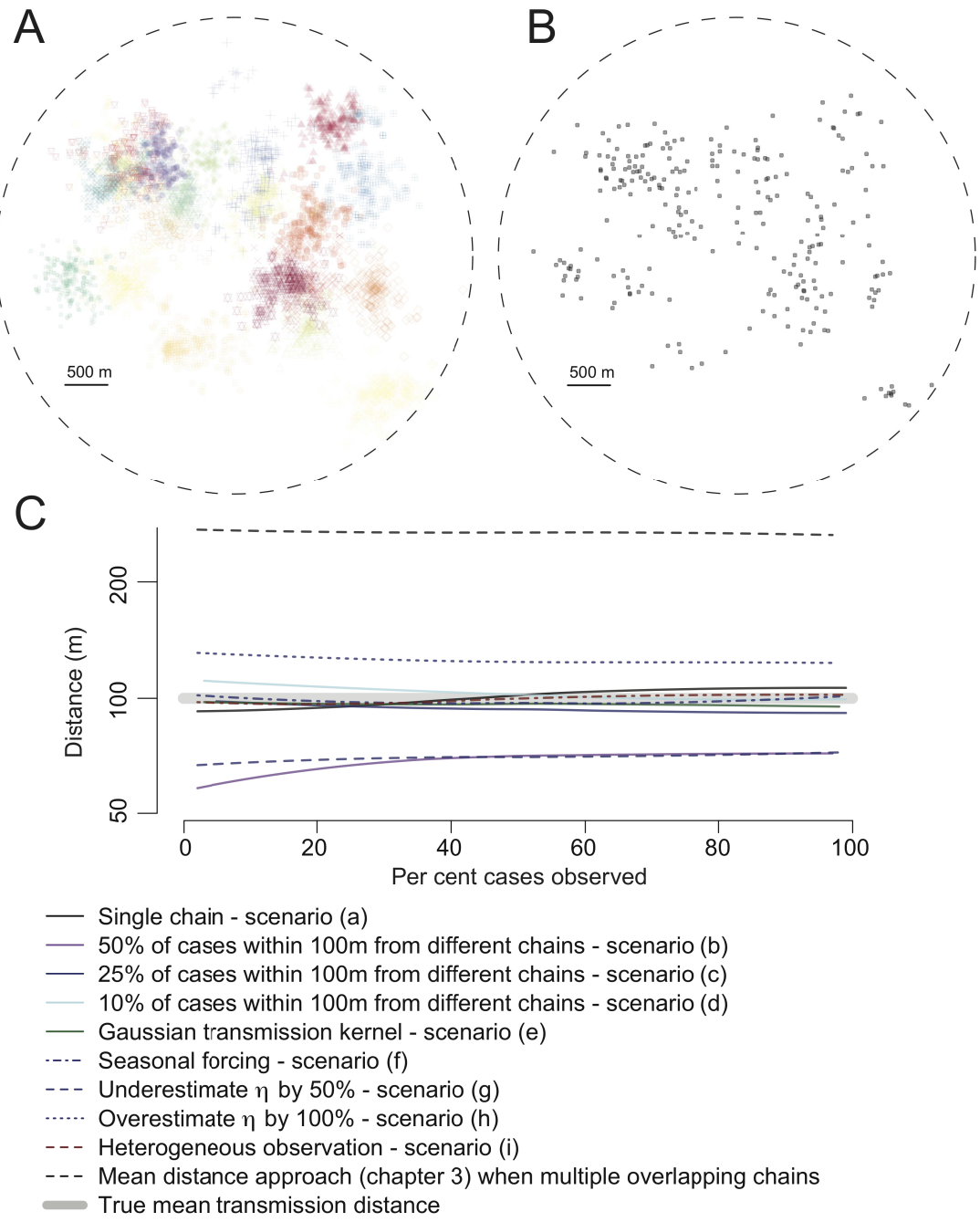


Figure 4.3: (A) Example map of simulated cases from 20 transmission chains with a mean transmission distance of 100m where all cases are observed (average 25% of cases within 100m are from different chains). Each color represents a different chain. (B) Example of the cases that are actually observed (5% of all cases) - note that in the analysis we cannot differentiate between the different chains. (C) Estimated mean transmission distance from the different scenarios listed in Table 4.1. Each line represents the loess curve from 500 simulations

4.4.2 Transmission distance of dengue in Bangkok

Having demonstrated the robustness of our approach using simulated data, we applied it to dengue data from Bangkok. A city that has experienced endemic dengue for decades. We were able to successfully geocode 8620 of the 11612 confirmed dengue patients (74%). Of these cases, 6305 lived within 10km of the hospital (73% of successfully geocoded cases) (Figure 4.4). Serotype data was available for 58% of cases.

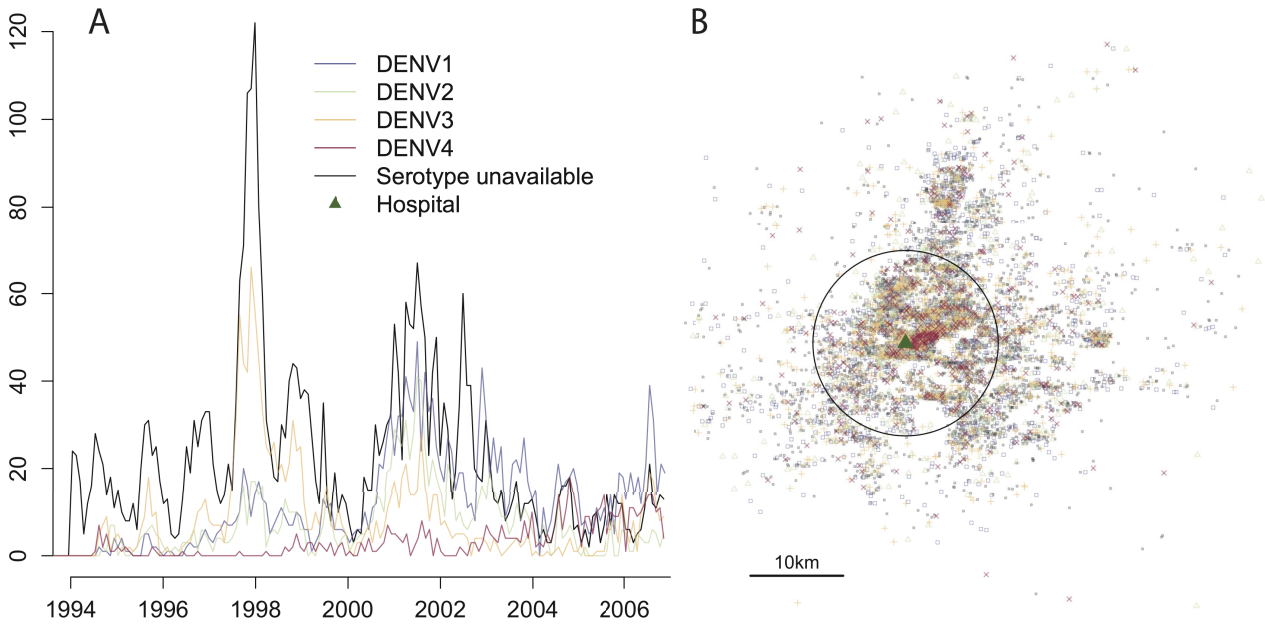


Figure 4.4: Spatial and temporal distribution of 8,620 dengue cases that presented at Queen Sirikit Hospital between 1994 and 2006. (A) Number of cases per month per serotype. (B) Location of patient homes. Only cases within the black circle were used in the estimation of the transmission distance.

The proportion of two cases occurring within a month of each other when separated by over 5km (P_0) that were homotypic was 0.37. The proportion of two cases

that were homotypic at 1km (P_x, the maximum pair-wise distance considered in our analysis) was 0.61 giving an estimate of 0.38 for $\hat{\eta}$ and an estimated number of circulating transmission chains within 1km of 2.6.

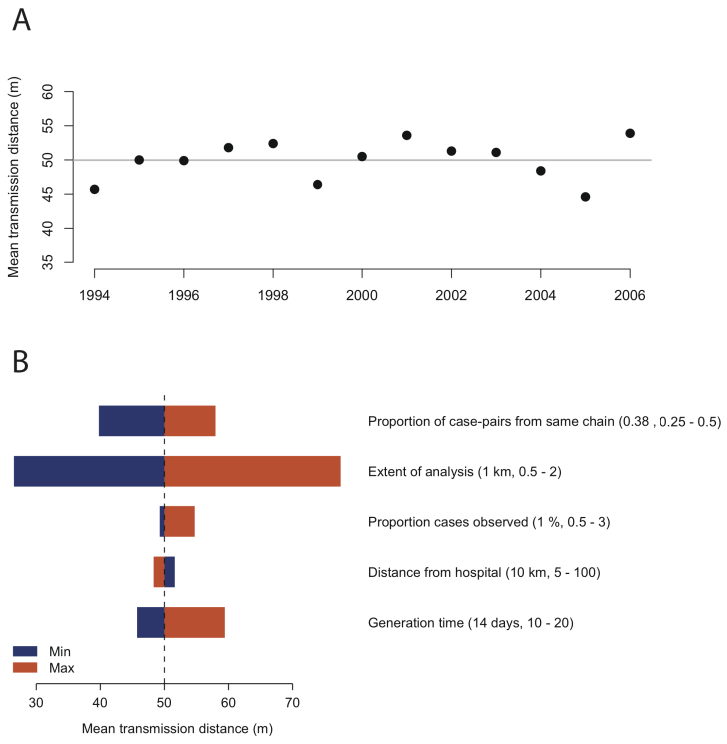


Figure 4.5: (A) Estimated mean transmission distance for dengue cases in Bangkok between 1994 and 2006. The grey line represents the overall mean of $50m$. (B) Sensitivity of overall mean to different model inputs.

Using this estimate of η, we calculated the mean transmission distance for each year between 1994 and 2006. We found a remarkably consistent pattern across the entire study period with an overall mean transmission distance of 50m, ranging from 44m to 54m within any year over the 13 year period, with no clear secular trend (Figure 4.5A). These findings indicate that the mean transmission distance is not much farther than the distance between neighboring houses.

Our estimates were robust to substantial difference in the parameter inputs (Fig-

ure 4.5B). They were most sensitive to the maximum distance at which we performed the analysis, however, doubling the distance (equivalent to quadrupling the area considered for each index case), resulted in only a 25m difference in the estimated mean transmission distance.

4.5 Discussion

Understanding the distance between sequential cases in a transmission chain is paramount to elucidating dispersal mechanisms and designing efficient intervention measures. However, characterizing transmission distances has been hampered by the presence of numerous overlapping transmission chains with no ability to differentiate between the different chains. In addition, we rarely observe all infections due to poor surveillance and the frequency at which individuals only suffer mild symptoms or are even completely unaware they have been infected [1]. To address this gap, we developed an approach to estimate the mean transmission distance between cases in the presence of numerous transmission chains and partially observed data. When fewer than five percent of cases were observed we able to accurately estimate the true transmission distance even in the presence of numerous overlapping transmission chains.

The distance between sequential dengue cases has been especially difficult to understand. The presence of an intermediary vector results in related infections in individuals that may never have been in contact. We applied our novel method to geocoded dengue cases from Bangkok, Thailand, a setting that experiences thousands of hospitalizations from the pathogen each year [2, 3]. We found a mean transmission distance of 50m, with virtually no differences by year, was most consistent with

the observed spatio-temporal pattern of cases. These findings are consistent with small scale transmission, not much greater than the distance between neighboring households, driving the spread of the disease. The dengue vector, *Aedes aegypti*, does not travel very far with mark, release and recapture experiments finding that the majority of mosquitoes will remain in the same or neighboring household upon release [12]. Human movement patterns also appear critical. A recent study in Iquitos, Peru found that previous detection of disease in an individual's 'Activity Space' (where people spend their time), was correlated with infection risk [13]. Activity Space closely related to home location. Individuals' movements during mosquito biting times (mornings and evenings) may largely drive the movement of the virus [14].

Using similar data from Bangkok, it has previously been demonstrated that there exists spatial dependence between cases of dengue occurring within a month of each other at distances up to 1km [15]. Our findings here indicate that small scale movements can generate a large footprint of spatial dependence. Cases occurring several hundreds of meters away at the same time may still originate from the same transmission chain but their infector in common may be several generations back.

The presented methods will be applicable across disease symptoms. It will help understand how diseases move around even where we are only able to observe a tiny proportion of all cases. This in turn will inform spatially explicit modeling efforts that attempt to understand the potential impact of interventions, including insecticide spraying and vaccines.

The requirement to know the number of circulating transmission chains maybe be difficult to estimate, however, we have demonstrated how we can use smaller areas where estimates may be available. In addition, our estimate were found to be robust to even significant misspecification of the true number of circulating chains.

We have assumed that the mean and standard deviation of the transmission kernel are the same, such as is found with the exponential distribution. Furthermore the presented method can clearly be adjusted to identify the combinations of the standard deviation and mean of the transmission kernel that are most consistent with the data (although as demonstrated in Chapter 3, a range of similar combinations will be equally consistent). In settings of extreme overlap between transmission chains where more than half of closest cases come from different transmission chains, our approach will underestimate the true transmission distance. While this appears unlikely, genetic approaches may help our understanding of the number and overlap between different transmission chains within any area. Finally, our approach requires that the central limit theorem applies after a small number of generations. Where the transmission kernel is best approximated by long-tailed distributions (such as the Pareto), the estimates from our approach will represent the truncated form of the distribution so that it will not be able to identify the tail of the distribution, however, it will capture the mean distance of the bulk of the transmissions.

In conclusion, understanding transmission distances is critical to elucidating pathogen dispersal mechanisms and designing interventions. The present approach will help in this regard across disease systems.

References

[1] S. B. Halstead, "Dengue," Imperial College Press, London, Oct. 2008.

[2] A. Nisalak, T. P. Endy, S. Nimmannitya, S. Kalayanarooj, U. Thisayakorn, R. M. Scott, D. S. Burke, C. H. Hoke, B. L. Innis, and D. W. Vaughn, "Serotype-specific dengue virus circulation and dengue disease in Bangkok, Thailand from 1973 to 1999." *The American journal of tropical medicine and hygiene*, vol. 68, no. 2, pp. 191–202, Feb. 2003.

[3] D. S. Burke, A. Nisalak, D. E. Johnson, and R. M. Scott, "A prospective study of dengue infections in Bangkok." *The American journal of tropical medicine and hygiene*, vol. 38, no. 1, pp. 172–180, Jan. 1988.

[4] D. J. Gubler, "Dengue and dengue hemorrhagic fever." *Clinical microbiology reviews*, vol. 11, no. 3, pp. 480–496, Jul. 1998.

[5] A. A. Sabchareon, D. D. Wallace, C. C. Sirivichayakul, K. K. Limkittikul, P. P. Chanthavanich, S. S. Suvannadabba, V. V. Jiwariyavej, W. W. Dulyachai, K. K. Pengsaa, T. A. T. Wartel, A. A. Moureau, M. M. Saville, A. A. Bouckenooghe, S. S. Viviani, N. G. N. Tornieporth, and J. J. Lang, "Protective efficacy of the recombinant, live-attenuated, CYD tetravalent dengue vaccine in Thai schoolchil-

dren: a randomised, controlled phase 2b trial." *The Lancet*, vol. 380, no. 9853, pp. 1559–1567, Nov. 2012.

[6] E. A. Codling, M. J. Plank, and S. Benhamou, "Random walk models in biology." *Journal of the Royal Society, Interface / the Royal Society*, vol. 5, no. 25, pp. 813–834, Aug. 2008.

[7] J. Lieblein, "On Moments of Order Statistics from the Weibull Distribution," *The Annals of Mathematical Statistics*, vol. 26, no. 2, pp. 330–333, Jun. 1955.

[8] E. C. Holmes, L. M. Bartley, and G. P. Garnett, "10 The Emergence of Dengue: Past Present and Future," in *Biomedical Research Reports*. Elsevier, 1998, pp. 301–325.

[9] N. G. Reich, S. Shrestha, A. A. King, P. Rohani, J. Lessler, S. Kalayanarooj, I.-K. Yoon, R. V. Gibbons, D. S. Burke, and D. A. T. Cummings, "Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity." *Journal of the Royal Society, Interface / the Royal Society*, vol. 10, no. 86, pp. 20 130 414–20 130 414, Jan. 2013.

[10] J. Aldstadt, "An incremental Knox test for the determination of the serial interval between successive cases of an infectious disease," *Stochastic Environmental Research and Risk Assessment*, vol. 21, no. 5, pp. 487–500, Aug. 2007.

[11] J. Aldstadt, I.-K. Yoon, D. Tannitisupawong, R. G. Jarman, S. J. Thomas, R. V. Gibbons, A. Uppapong, S. Iamsirithaworn, A. L. Rothman, T. W. Scott, and T. Endy, "Space-time analysis of hospitalised dengue patients in rural Thailand reveals important temporal intervals in the pattern of dengue virus transmission." *Tropical medicine & international health : TM & IH*, Jul. 2012.

[12] L. C. Harrington, T. W. Scott, K. Lerdthusnee, R. C. Coleman, A. Costero, G. G. Clark, J. J. Jones, S. Kitthawee, P. Kittayapong, R. Sithiprasasna, and J. D. Edman, "Dispersal of the dengue vector Aedes aegypti within and between rural communities." *The American journal of tropical medicine and hygiene*, vol. 72, no. 2, pp. 209–220, Feb. 2005.

[13] S. T. Stoddard, B. M. Forshey, A. C. Morrison, V. A. Paz-Soldan, G. M. Vazquez-Prokopec, H. Astete, R. C. Reiner, S. Vilcarromero, J. P. Elder, E. S. Halsey, T. J. Kochel, U. Kitron, and T. W. Scott, "House-to-house human movement drives dengue virus transmission," *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

[14] M. Yasuno and R. J. Tonn, "A study of biting habits of Aedes aegypti in Bangkok, Thailand." *Bulletin of the World Health Organization*, vol. 43, no. 2, pp. 319–325, 1970.

[15] H. Salje, J. Lessler, T. P. Endy, F. C. Curriero, R. V. Gibbons, A. Nisalak, S. Nimmannitya, S. Kalayanarooj, R. G. Jarman, S. J. Thomas, D. S. Burke, and D. A. T. Cummings, "Revealing the microscale spatial signature of dengue transmission and immunity in an urban population." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 24, pp. 9535–9538, Jun. 2012.

Part II

Neutralization titer considerations

Chapter 5

Characterizing the variability of the dengue Plaque Reduction Neutralization Assay

Henrik Salje,Isabel Rodrguez-Barraquer, Kaitlin Rainwater-Lovett, Ananda Nisalak, Butsaya Thaisomboonsuk, Stephen J. Thomas, Stefan Fernandez, Richard G. Jarman, In-Kyu Yoon and Derek A. T. Cummings

5.1 Abstract

Accurate determination of neutralization antibody titers supports epidemiological studies of dengue virus transmission and vaccine trials. Neutralization titers measured using the plaque reduction neutralization test (PRNT) is believed to provide a key measure of immunity to dengue viruses, however, the assays variability is poorly understood, making it difficult to understand the significance of any assay reading. In addition there is limited standardization of the PRNT cut-point or statistical model used to estimate titers across laboratories, with little understanding of the optimum approach. We used repeated assays on the same two pools of serum using five different viruses (2,319 assays) to characterize the variability in the technique under identical experimental conditions. We also assessed the performance of multiple

statistical models to interpolate continuous values of neutralization titer from discrete measurements from serial dilutions and identified the optimal PRNT cut-point for the assay. We found that titer estimates varied widely with an average standard deviation of 0.18 (logarithmic scale). We estimate that for a true $PRNT_{50}$ titer of 1:300, 95% of measured titers will range from 1:140 to 1:720. Further, we found that a common statistical model, probit regression, consistently overestimated titers whereas the alternative cloglog regression and four-parameter non-linear regression were largely unbiased. Finally, optimum PRNT cut-points ranged from $PRNT_{65}$ to $PRNT_{75}$, depending on the statistical model used. Researchers should consider PRNT variability when characterizing an individuals immune status.

5.2 Introduction

Dengue remains a substantial public health problem in tropical and subtropical regions [1]. All four serotypes of the mosquito-borne virus are capable of producing significant morbidity and death [2]. As part of efforts to monitor and control the disease, public health agencies and vaccine developers use serological methods to perform surveillance and assess vaccine trial outcomes. A standard for characterizing serotype-specific neutralizing dengue antibody levels is the Plaque Reduction Neutralization Test (PRNT) [3]. PRNT readouts are known to vary substantially, even on samples from the same individual, however, the extent of the underlying variability in estimates remains unclear [4]. There are many potential sources of variation including within experiment and between experiment sources. In addition, different laboratories use different cell lines, different viral strains with varying viral passage number, and parametric models to calculate PRNT with the impact of the alter-

native approaches poorly understood [5–7]. Laboratories also use PRNT cut-points that range between $PRNT_{50}$ to $PRNT_{90}$, and may perform varying numbers of serial dilutions [6, 8–10]. Understanding and characterizing the variability of the assay may greatly increase the accuracy and quantifiability of the assay, important both in epidemiological and vaccine studies.

After infection by one of the four dengue virus serotypes, individuals develop antibodies against the infecting virus [2]. The PRNT assay is used to measure neutralizing antibodies produced in response to this exposure. When an *in vitro* monolayer of cells is exposed to the virus without the presence of neutralizing antibodies, the viral particles enter and kill the cells. Where viral particles have spread between neighboring cells, a plaque of dead cells is created that can be observed and counted. The presence of neutralizing antibodies from an individuals serum reduces the number of plaques formed by inhibiting the virus. In most cases, for a given concentration of antibodies, the addition of lower dilutions of serum result in fewer plaques formed than higher serum dilutions. $PRNT_{50}$ is the estimated serum dilution that produces a 50% reduction in the number of plaques formed compared to the number formed on monolayers in the absence of antibody [3]. $PRNT_{50}$ is believed to give an indication of an individuals ability to neutralize the dengue virus if exposed in vivo and to indicate whether an individual has been exposed in the past.

An individuals ability to successfully neutralize a strain of dengue may depend on the age of the individual, gender, nutrition, genetic factors as well as the history and time of previous infections by other flaviviruses [2, 11]. In comparing single PRNT estimates between individuals, it is not possible to separate differences due to these host factors from differences due to assay variability. Understanding the variability of the assay instead requires a large number of repeated experiments on the same

serum. This necessitates large pools of serum that are rarely available. However, as part of each experiment, laboratories often use high titer and low titer serum controls to ensure consistency of experimental conditions between assays. Control sera lots can come from pooled human sera that are maintained and remain unchanged for several years. In each experiment, PRNTs are calculated for each control serum (as well as the test serum under investigation). Using the plaque counts from the control sera from a large number of assays, we can estimate the variability in the PRNT within identical experiments.

5.3 Methods

The Armed Forces Research Institute of Medical Sciences (AFRIMS) in Bangkok, Thailand developed the dengue PRNT assay in the 1960s and has been performing it since for surveillance of dengue immunity in the population and supporting vaccine trials and cohort studies [3, 12, 13]. Data for the current study comes from control assays of PRNTs performed at AFRIMS between 2007 and 2013. Briefly, in each assay, a monolayer of Macaca mulatta kidney cells (LLC-MK2) was infected with virus, predetermined to be in the range of 30-50 plaque-forming units in the presence of 4-fold serial dilutions of heat-inactivated serum (range of 1:10 to 1: 163840). For each dilution, the number of viral plaques was counted and compared to the number of plaques in a control where no serum was added. Each dilution and control was performed in duplicate and the plaque count averaged across the two repeats. During the study period there were changes to the number and cell lines used to passage the virus, and the number of passages that the virus went through. In addition, the DENV-4 viral strain was changed in 2009 (Table 5.1). Three technicians conducted

over 95% of all assays in the study period.

5.3.1 Serum pools

Two serum pools (a high titer and a low titer pool) were collected and created in 2006 and used throughout the study period. The high titer pool was obtained by pooling residual blood samples from multiple Thai individuals that tested positive for dengue virus using IgG ELISA. A portion of the pool was then diluted with human sera from PRNT-negative blood donors to create a low titer pool.

5.3.2 Viruses

Five viruses were used during the study period, one each for DENV-1, DENV-2 and DENV-3 and two for DENV-4 (Table 5.1). Around every two years, viral stocks were generated in batches by passaging virus through C6/36 mosquito cell lines (between one and eight passages) and up to three passages in either suckling mice (SM) or LLC-MK2 cells.

Serotype	Serum pool	Neutraliziing strain	N (%)	$PRNT_{50}$
DENV- 1	High titer	Thailand/16007/1964	288 (12)	1:10700
DENV- 1	Low titer	Thailand/16007/1964	288 (12)	1:1200
DENV- 2	High titer	Thailand/16681/1984	279 (12)	1:12900
DENV- 2	Low titer	Thailand/16681/1984	279 (12)	1:1400
DENV- 3	High titer	Philippines/16562/1964	285 (12)	1:7500
DENV- 3	Low titer	Philippines/16562/1964	285 (12)	1:800
DENV- 4A	High titer	Indonesia/1036/1976	179 (8)	1:500
DENV- 4A	Low titer	Indonesia/1036/1976	180 (8)	1:50
DENV- 4B	High titer	Thailand/C0036/2006	128 (5)	1:5600
DENV- 4B	Low titer	Thailand/C0036/2006	128 (5)	1:600

Table 5.1: Number of experiments by serum pool and viral strain combination. The final column shows the $PRNT_{50}$ calculated using a smooth spline from all experiments.

5.3.3 PRNT calculation

Basic regressions were used to interpolate the titer at which defined reductions (PRNT cut-points) occur from the observed reductions (e.g. a 50% reduction for a cut-point of $PRNT_{50}$). We calculated PRNTs over the range $PRNT_{40}$ to $PRNT_{90}$ using either (a) probit regression, (b) logistic regression, (c) complementary log-log (cloglog) regression or (d) four-parameter non-linear regression [8].

As PRNTs can be resource intensive, laboratories may perform two dilutions that they expect will contain the PRNT cut-point of interest and use straight line interpolation on the log-transformed dilutions [6]. To estimate the variability of this approach, we initially identified the expected PRNT titer using all assays from a viral strain

and serum pool and identified the two sequential dilutions that contained this value. For each experiment we then only used the values from the two sequential dilutions to calculate PRNT using straight-line interpolation. We did not calculate PRNTs in experiments where the two dilutions did not contain the cut-point of interest.

Finally, some laboratories only perform a single dilution and calculate the neutralization titer at that single dilution (known as a Single Dilution Neutralization Test, SDNT) [14]. To estimate the variability in SDNTs, we calculated the variance in the neutralization proportions for all experiments from each individual dilution for each viral strain and serum pool.

5.3.4 Bias and mean squared error

We assumed that the ability of each of the two serum pools to neutralize a particular viral strain was constant, reflected in a single true PRNT titer for each virus for both the high titer and the lower titer pools (i.e., one for each row in Table 5.1). We considered PRNT estimates from a flexible non-parametric spline, fitted to the plaque reductions from all experiments from a single serum pool as the best, unbiased estimate of the true PRNT for that pool.

We explored whether there existed any systematic differences (bias) in PRNT estimates calculated using the different models. For each experiment, we calculated PRNT titers using each of the models (probit, logit, cloglog regression and non-linear regression). Bias was suggested when there was a systematic difference between the PRNT estimates using the model and the estimate of the true titer. In addition we calculated the mean squared error (MSE) in the estimates. We reported an average MSE, bias and variance for each cut-point and model, weighted by the number of

experiments using each virus and serum pool.

Detailed methods of the PRNT calculation, estimation of bias and variance and the calculation of the confidence intervals can be found in the supplementary materials.

5.3.5 Multilevel model

Heterogeneities in the passaging of the virus may be associated with changing PRNT estimates. To quantify systematic differences in $PRNT_{50}$ estimates by the number of passages and the type of cell (C6/36, SM and LLC-MK2 cells) and the age of the virus stock, we constructed a multilevel model with a random intercept for each viral strain and serum pool combination (listed in Table 5.1).

5.3.6 Ethics statement

All experiments were conducted using pooled residual sera from public health service testing and, as per Walter Read Armed Institute of Research (WRAIR) policy, did not require ethics review. WRAIR is the parent organization of AFRIMS.

5.4 Results

Between 2007 and 2013, a total of 2,319 PRNTs were performed using five different viruses on two different control sera (Table 5.1). There existed substantial variability in the plaque reduction proportions (Figure 5.1). On average, the plaque reduction proportions had a standard deviation of 0.10 within each dilution. These findings

indicate that an estimated SDNT of 70% has a 95% confidence interval of 50% - 90% (Figure 5.2).

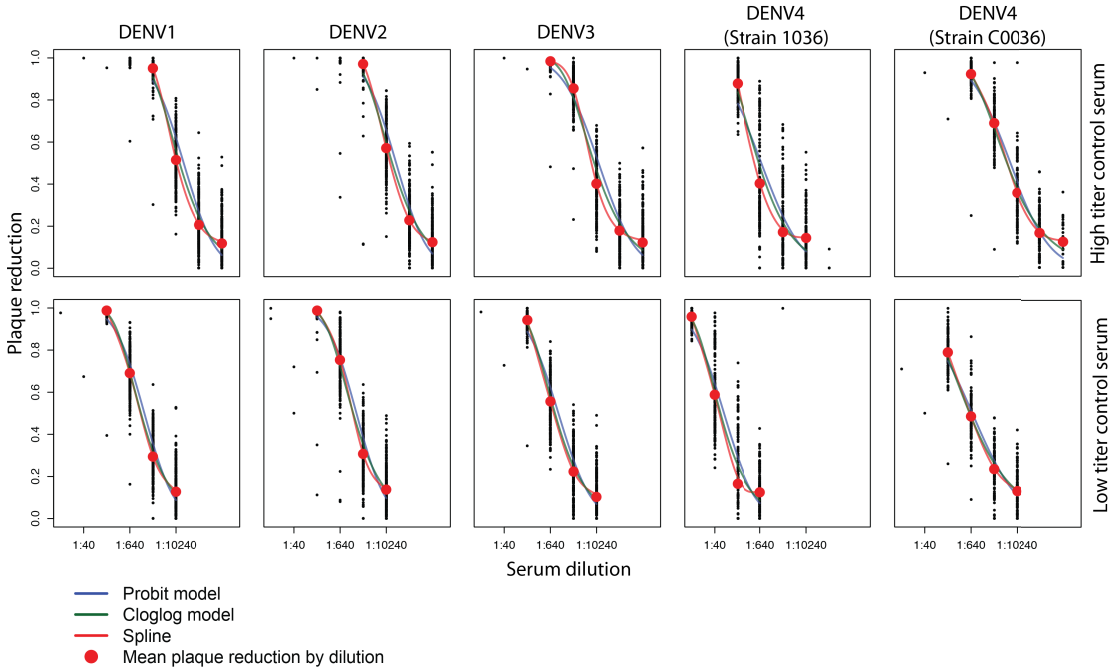


Figure 5.1: Plaque reduction estimates for each experiment. Each black dot represents the mean reduction in plaques formed for that dilution from two repeats. The red dots are the overall means across all the experiments. Superimposed are fitted models using a probit transformation, a cloglog transformation and a non-parametric spline.

The variability in plaque reduction proportions led to heterogeneity in PRNT titer estimates (Figure 5.2). $PRNT_{50}$ titer estimates using the different models had similar levels of variability: the conventional probit model had a mean standard deviation of 0.20 (log10 scale, range of 0.15 - 0.28 from the ten different serum pool - viral strain combinations listed in Table 5.1), the logit model had a mean standard deviation of

0.20 (0.15 - 0.29), the cloglog model had a mean standard deviation of 0.20 (0.15 - 0.27) and four-parameter non-linear regression had a mean standard deviation of 0.18 (0.15 - 0.28) (Table 5.2).

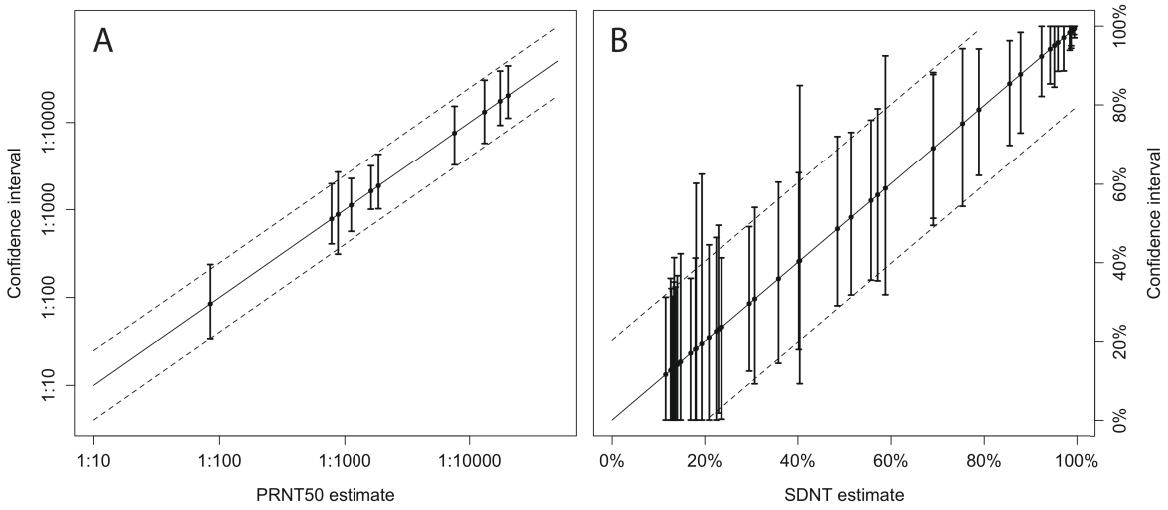


Figure 5.2: Confidence intervals for $PRNT_{50}$ and SDNT. (A) The solid lines represents 95% confidence intervals using the 2.5% and 97.5% quantiles from all experiments from each of the ten serum pools. The dotted lines represent asymptotic 95% confidence intervals using the standard deviation from all experiments using probit regression. (B) The solid lines represents 95% confidence intervals using the 2.5% and 97.5% quantiles from each dilution and serum pool. The dotted lines represent asymptotic 95% confidence intervals using the standard deviation from all dilutions.

We found that the probit and logit models consistently overestimated $PRNT_{50}$, by an average of 0.14 (log10 scale) and 0.12, respectively. The cloglog and four-parameter non-linear regression approaches by contrast were largely unbiased. As an example of what this translates to on a linear scale: an individual with a true $PRNT_{50}$ titer of 1:300 would have a mean measured $PRNT_{50}$ titer of 1:410 with a 95% confidence interval of 1:170 to 1:1020 when using a probit model, a mean measured $PRNT_{50}$ titer of 1:340 with a 95% confidence interval of 1:140 to 1:830 when using a cloglog model and a mean measured $PRNT_{50}$ titer of 1:320 with a 95% confidence interval

of 1:140 to 1:720 when using four-parameter non-linear regression. The extent of the bias using probit models appeared to be constant across titers (log scale, Figure C.1). Ratios of PRNT estimates are used in the detection of seroconversion (for example in the comparison of pre and post infection serum). While individual PRNT estimates may be biased using probit transformations, ratios of PRNTs would not be as both the numerator and the denominator would be similarly biased.

Model	Standard deviation [range]	Bias [range]
Probit regression	0.20 [0.15 - 0.29]	0.14 [0.07 - 0.19]
Logistic regression	0.20 [0.15 - 0.29]	0.12 [0.06 - 0.18]
Cloglog regression	0.20 [0.15 - 0.27]	0.05 [-0.01 - 0.12]
Four-parameter non-linear regression	0.18 [0.15 - 0.28]	0.02 [-0.02 - 0.32]

Table 5.2: Standard deviation and bias in $PRNT_{50}$ estimates using the different models. Reported values are average from the different viruses and serum pools, weighted by the number of experiments.

Some laboratories use different PRNT cut-points (e.g., $PRNT_{50}$ vs. $PRNT_{90}$). The MSE for each model can be used to discriminate the performance of models using different cut-points. The lowest MSE for the probit, logit and cloglog models were at $PRNT_{75}$ and at $PRNT_{65}$ for four-parameter non-linear regression (Figure 5.3). The MSE for the four-parameter non-linear regression was the lowest among the different models for cut-points below $PRNT_{65}$, whereas the cloglog model was lowest for higher cut-points.

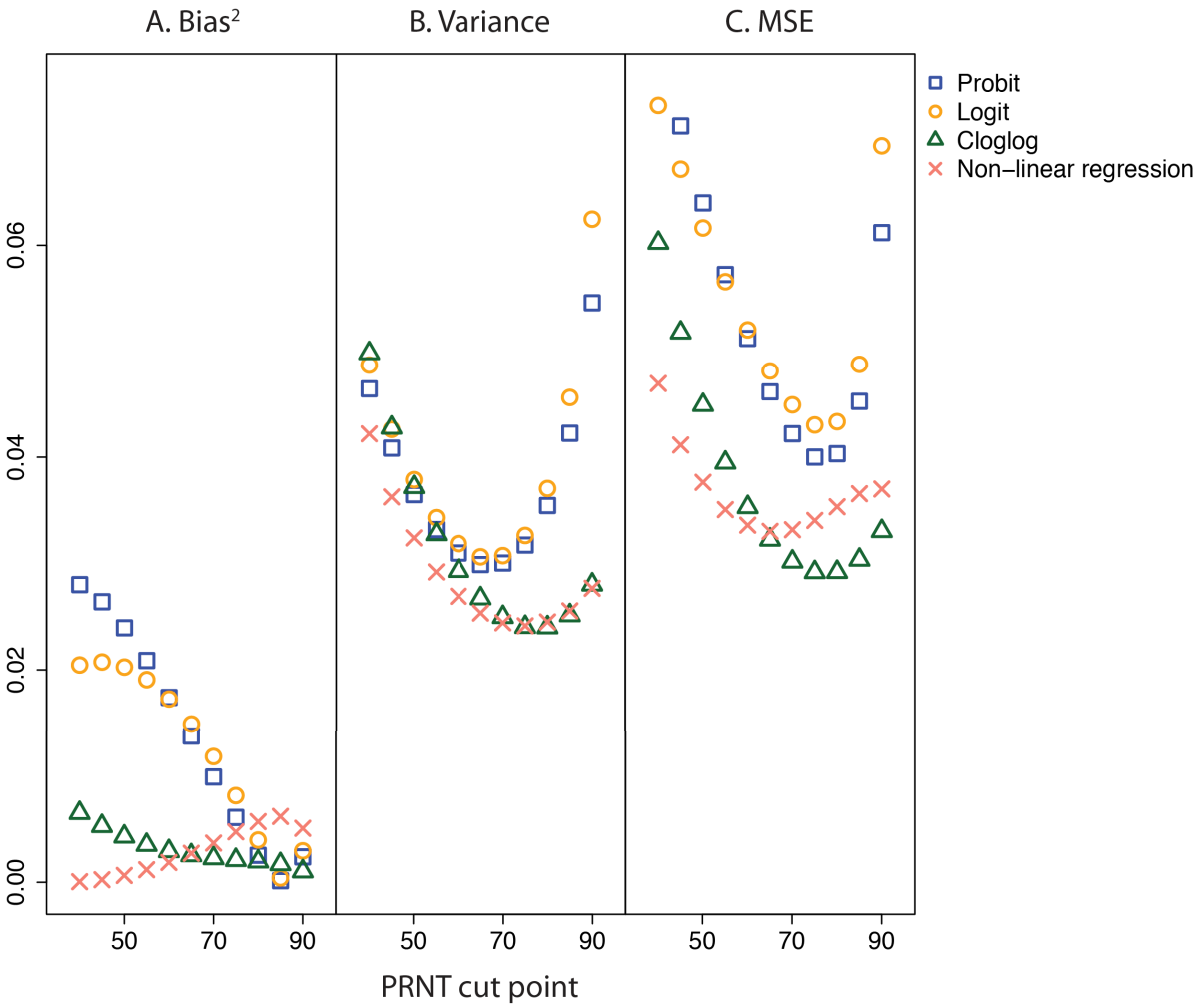


Figure 5.3: Estimates of (A) bias, (B) variance and (C) mean squared error by PRNT cut-point for the different models. Bias for each experiment was calculated by comparing the model PRNT results with that from a smooth spline from all experiments from that particular virus and serum pool.

Where only two dilutions were used, only 50% of the experiments could be used as the two sequential dilutions did not contain the PRNT cut-point in the remainder and would have required extrapolation. Where it could be estimated, the standard deviation of $PRNT_{50}$ using two dilutions was estimated at 0.13 (range: 0.09 - 0.18), however, this only represents the variability of the subset of the experiments where the two dilutions had reductions in plaques that were closest to the best estimate of

the unbiased PRNT.

To estimate the effects of experimental conditions on $PRNT_{50}$ titers we built a multilevel model incorporating the number of viral passages, the cell type and the age of the virus stock used in the experiments. We found that passaging the virus in SM increased titers compared to LLC-MK2 cells (effect size of 1.17, 95% confidence interval of 1.11 - 1.25). The total number of passages and the age of the viral stock at the time of the experiment did not affect the PRNT titers. Less than 0.2% of the variability in $PRNT_{50}$ estimates could be explained by the model covariates, leaving over 99% of variability unexplained (Table 5.3).

Parameter		Coefficient [95% CI]
Age of virus stock (yrs):	mean: 2.2, sd: 1.0	0.98 [0.96 1.00]
Total # of passages:	mean: 5.3, sd: 2.1	1.01 [1.00 1.02]
Cell passage:		
- C6/36 and LLC-MK2	# experiments: 502	Ref
- C6/36 and SM	# experiments: 1566	1.17 [1.11 1.25]
- Only C6/36	# experiments: 262	1.01 [0.06 19.9]
R^2 (1)		0.001

Table 5.3: Results of multilevel model for impact of experimental factors on $PRNT_{50}$ estimates using a probit transformation. The model has a random intercept for the viral strain and serum pool used in the experiment. All coefficients have been transformed by raising 10 to the power of the coefficient. (1) Marginal R^2 that indicates the proportion of variance explained by the fixed effects only [15].

5.5 Discussion

Using repeated assays on the same serum sample with the same viral strain, we estimated the extent to which measured PRNTs vary. We found a consistent level of variability in titer estimates across the viruses and serum pools used during the study period. Our findings indicate that if, for example, a $PRNT_{50}$ titer of 1:300 were to be considered a true surrogate of protection, researchers should only consider measured $PRNT_{50}$ titers of 1:750 or greater as strong evidence of an individuals titers being sufficiently high. A measure of the variability of $PRNT_{50}$ results provides information on the potential misclassification of individuals falling above or below any specified cut-point, information routinely used in calculating sample sizes for a wide range of studies. By characterizing the variability in measured titers, these findings will aid in the determination of an individuals immunity, the design and interpretation of results from immunogenicity trials, epidemiologic studies and allow the benchmarking of assays across laboratories.

Despite efforts to standardize the assay, heterogeneities in approaches between laboratories persist [7, 16, 17]. In particular different PRNT cut-points are common. The WHO recommends using a $PRNT_{50}$ titer for vaccinee sera and $PRNT_{90}$ titers for epidemiological studies and diagnosis. The stated benefits of the higher cut-point is to decrease both variability in the estimates and to minimize the effects of cross-reaction from other flaviviruses such as other dengue serotypes and Japanese Encephalitis, although the extent to which this occurs remains to be fully understood [17]. Lower cut-points improve the sensitivity of the assay at the expense of increasing the risk of falsely classifying susceptible individuals as protected. Research studies often use cut-points between $PRNT_{50}$ and $PRNT_{90}$ [6, 10]. We found that the cut-

point that minimized the MSE between the model PRNT estimates and our best estimate of the unbiased PRNTs as between $PRNT_{65}$ and $PRNT_{75}$. Studies should consider cut-points in this range where assay sensitivity and specificity requirements are met. Using four-parameter non-linear regression (e.g. as implemented in Prism software, specifically mentioned in the WHO guidelines) typically requires at least four serum dilutions. A popular alternative method that requires fewer dilutions is probit regression. We found this method produced substantially biased PRNT estimates across all cut-points. Cloglog regression was largely unbiased. Overall the four-parameter non-linear regression performed best for cut-points below $PRNT_{65}$, whereas the cloglog model was best for higher cut-points.

Laboratories may perform only two dilutions and use linear interpolation to obtain PRNT estimates. We found that we could only use 50% of the assays for this analysis, as the remaining experiments would require unwise extrapolation outside the results from the two dilutions. In these situations, laboratories need to repeat the assays at different dilution ranges. The substantial number of experiments that could not be included in the analysis suggests that performing only two dilutions may only have minimal benefits. Single dilution neutralization tests only require a single dilution, however, we found that SDNT estimates had wide confidence intervals: we estimated that a particular dilution of sera that had a true SDNT result of 70%, the typical cut-point used in epidemiological studies, had 95% confidence intervals of between 50% and 90% [14].

The viral strain used in the assay has been suggested to cause systematic differences [7]. The DENV-4 strain used in the assays was changed in 2009 resulting in a 11.4-fold increase in mean $PRNT_{50}$ titers in the high serum pool and a 12.1-fold increase in the low serum pool, confirming previous results [4]. A potential expla-

nation for this substantial difference in titer is the evolution of the virus between 1976 (the date of the original virus) to the newer 2006 virus, resulting in immunologically different responses. Alongside the effect of viral strain, it has been suggested that the number and cell type of viral passages could produce systematic differences in PRNT estimates [4, 7]. We found a small increase in titers in experiments using viruses passaged through SM compared to LLC-MK2 cells supporting similar previous findings [4]. The total number of viral passages did not appear to impact PRNT estimates, however, only small numbers of passages were conducted (maximum of eight). Increasing this substantially may nevertheless impact estimates. Overall, aside from viral strain, experimental factors explained less than one per cent of the variability in PRNT estimates.

Our findings suggest that the assay is inherently variable. There are many potential sources of variability in each experiment: (a) the number of viral particles pipetted into each plate, (b) the extent of viral - antibody interaction (c) the spatial arrangement of cells in the monolayer and (d) the number of non-overlapping plaques successfully generated and counted. While technicians can minimize differences through effective mixing and careful dilutions, there may be a limit to the extent that variability in these factors can be reduced. The use of automated counting methods that allow faster and more accurate particle counting may help [18]. In addition, the use of Reporter Virus Particles in laboratories with access to flow cytometry equipment and microneutralization assays show some encouraging results, although further work is needed to quantify their variability [18–20].

This study demonstrates the utility of raw results. Laboratories should consider reporting plaque counts alongside titer estimates. This will allow investigators to easily compute alternative titers using different cut-points or statistical models, facil-

itating comparison across laboratories.

The study had some limitations. The serum pools come from pooled human sera that contain a wide range of antibodies not representative of a single individuals serum. Nevertheless, the ability for the pooled serum to neutralize a single virus should remain constant. Further, serum with neutralization titers outside the range used in this study may perform differently. The range of titers in this study was wide ($PRNT_{50}$ range of 1:50 - 1:12900) and we observed a consistent pattern in variability across this range.

In conclusion, laboratories should consider the variability in the PRNT assay when characterizing the immunity of an individual. Where sufficient dilutions are performed, four-parameter non-linear regression should be used for PRNT cut-points less than $PRNT_{65}$ otherwise cloglog regression appears optimal.

References

[1] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, and S. I. Hay, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, Apr. 2013.

[2] S. B. Halstead, "Dengue," Imperial College Press, London, Oct. 2008.

[3] P. K. Russell, A. Nisalak, P. Sukhavachana, and S. Vivona, "A Plaque Reduction Test for Dengue Virus Neutralizing Antibodies," *The Journal of Immunology*, vol. 99, no. 2, pp. 285–290, Aug. 1967.

[4] S. J. Thomas, A. Nisalak, K. B. Anderson, D. H. Libraty, S. Kalayanarooj, D. W. Vaughn, R. Putnak, R. V. Gibbons, R. Jarman, and T. P. Endy, "Dengue plaque reduction neutralization test (PRNT) in primary and secondary dengue virus infections: How alterations in assay conditions impact performance." *The American journal of tropical medicine and hygiene*, vol. 81, no. 5, pp. 825–833, Nov. 2009.

[5] D. M. D. Morens, S. B. S. Halstead, P. M. P. Repik, R. R. Putvatana, and N. N. Raybourne, "Simplified plaque reduction neutralization assay for dengue viruses

by semimicro methods in BHK-21 cells: comparison of the BHK suspension test with standard plaque reduction neutralization." *Journal of Clinical Microbiology*, vol. 22, no. 2, pp. 250–254, Aug. 1985.

[6] A. C. Morrison, S. L. Minnick, C. Rocha, B. M. Forshey, S. T. Stoddard, A. Getis, D. A. Focks, K. L. Russell, J. G. Olson, P. J. Blair, D. M. Watts, M. Sihuincha, T. W. Scott, and T. J. Kochel, "Epidemiology of Dengue Virus in Iquitos, Peru 1999 to 2005: Interepidemic and Epidemic Patterns of Transmission," *PLoS Neglected Tropical Diseases*, vol. 4, no. 5, p. e670, May 2010.

[7] K. Rainwater-Lovett, I. Rodriguez-Barraquer, D. A. T. Cummings, and J. J. Lessler, "Variation in dengue virus plaque reduction neutralization testing: systematic review and pooled analysis." *BMC Infectious Diseases*, vol. 12, pp. 233–233, Jan. 2012.

[8] A. Puschnik, L. Lau, E. A. Cromwell, A. Balmaseda, S. Zompi, and E. Harris, "Correlation between Dengue-Specific Neutralizing Antibodies and Serum Avidity in Primary and Secondary Dengue Virus 3 Natural Infections in Humans." *PLoS Neglected Tropical Diseases*, vol. 7, no. 6, pp. e2274–e2274, Jun. 2013.

[9] K. Laoprasopwattana, D. H. Libraty, T. P. Endy, A. Nisalak, S. Chunsuttiwat, D. W. Vaughn, G. Reed, F. A. Ennis, A. L. Rothman, and S. Green, "Dengue Virus (DV) Enhancing Antibody Activity in Preillness Plasma Does Not Predict Subsequent Disease Severity or Viremia in Secondary DV Infection," *The Journal of infectious diseases*, vol. 192, pp. 510–519, Aug. 2005.

[10] J. H. McArthur, A. P. Durbin, J. A. Marron, K. A. Wanionek, B. Thumar, D. J. Pierro, A. C. Schmidt, J. E. Blaney, B. R. Murphy, and S. S. Whitehead,

"Phase I clinical evaluation of rDEN4Delta30-200,201: a live attenuated dengue 4 vaccine candidate designed for decreased hepatotoxicity." *The American journal of tropical medicine and hygiene*, vol. 79, no. 5, pp. 678–684, Nov. 2008.

[11] T. H. T. Nguyen, T. L. T. Nguyen, H.-Y. H. Lei, Y.-S. Y. Lin, B. L. B. Le, K.-J. K. Huang, C.-F. C. Lin, Q. H. Q. Do, T. Q. H. T. Vu, T. M. T. Lam, T.-M. T. Yeh, J.-H. J. Huang, C.-C. C. Liu, and S. B. S. Halstead, "Association between sex, nutritional status, severity of dengue hemorrhagic fever, and immune status in infants with dengue hemorrhagic fever." *The American journal of tropical medicine and hygiene*, vol. 72, no. 4, pp. 370–374, Apr. 2005.

[12] T. P. Endy, A. Nisalak, S. Chunsuttitwat, D. W. Vaughn, S. Green, F. A. Ennis, A. L. Rothman, and D. H. Libraty, "Relationship of Preexisting Dengue Virus (DV) Neutralizing Antibody Levels to Viremia and Severity of Disease in a Prospective Cohort Study of DV Infection in Thailand," *The Journal of infectious diseases*, vol. 189, pp. 990–1000, Mar. 2004.

[13] S. Simasathien, S. J. Thomas, V. Watanaveeradej, A. Nisalak, C. Barberousse, B. L. Innis, W. Sun, J. R. Putnak, K. H. Eckels, Y. Hutagalung, R. V. Gibbons, C. Zhang, R. De La Barrera, R. G. Jarman, W. Chawachalasai, and M. P. Mammen, "Safety and Immunogenicity of a Tetravalent Live-attenuated Dengue Vaccine in Flavivirus Naive Children," *The American journal of tropical medicine and hygiene*, vol. 78, no. 3, pp. 426–433, 2008.

[14] S. B. Halstead, "A prospective seroepidemiologic study on dengue in children four to nine years of age in Yogyakarta, Indonesia I. studies in 1995–1996," *The*

American journal of tropical medicine and hygiene, vol. 61, no. 3, pp. 412–419, 1999.

[15] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R2from generalized linear mixed-effects models," *Methods in Ecology and Evolution*, vol. 4, no. 2, pp. 133–142, Dec. 2012.

[16] J. T. J. Roehrig, J. J. Hombach, and A. D. T. A. Barrett, "Guidelines for Plaque-Reduction Neutralization Testing of Human Antibodies to Dengue Viruses." Jun. 2008.

[17] World Health Organization, "Guidelines for plaque-reduction neutralization testing of human antibodies to dengue viruses," *Immunization, Vaccines and Biologicals, World Health Organization*, Jul. 2007.

[18] W. W. S. I. Rodrigo, D. C. Alcena, R. C. Rose, X. Jin, and J. J. Schlesinger, "An automated Dengue virus microneutralization plaque assay performed in human Fcgamma receptor-expressing CV-1 cells." *The American journal of tropical medicine and hygiene*, vol. 80, no. 1, pp. 61–65, Jan. 2009.

[19] K. Mattia, B. A. Puffer, K. L. Williams, R. Gonzalez, M. Murray, E. Sluzas, D. Pagano, S. Ajith, M. Bower, E. Berdougo, E. Harris, and B. J. Doranz, "Dengue reporter virus particles for measuring neutralizing antibodies against each of the four dengue serotypes." *PloS one*, vol. 6, no. 11, p. e27252, 2011.

[20] C. C. Ansarah-Sobrinho, S. S. Nelson, C. A. C. Jost, S. S. S. Whitehead, and T. C. T. Pierson, "Temperature-dependent production of pseudoinfectious dengue reporter virus particles by complementation," *Virology*, vol. 381, no. 1, pp. 8–8, Nov. 2008.

Chapter 6

Conclusions

Characterizing the spatial dynamics of endemic diseases is arguably more difficult than describing the spread of newly emergent pathogens despite the volume of available data being far greater for the former. The presence of many overlapping transmission chains circulating in the same area for many years can be a significant hindrance to understanding how a pathogen is moving around a community. There is often no way to identify transmission-related case-pairs, especially when many infections only cause mild symptoms and are therefore not detected. In addition, where an intermediary vector exists, transmission-related individuals may have never come into contact with each other. The spatial analysis of dengue transmission suffers from all of these factors. Dengue has circulated in Southeast Asia for decades and in some urban centers such as Bangkok, endemic circulation of the virus never stops (albeit with some seasonal differences in the force of infection) [1]. There have been few previous attempts to characterize the small-scale spatial dynamics of the virus.

Here we developed novel approaches to characterize the spatial signature of infectious diseases and applied them to dengue and chikungunya. In particular we described approaches to characterize the global spatial dependence of infectious diseases (i.e., the tendency for cases to be found near each other) in scenarios where

there is information on the infecting pathogen (such as serotype). Using case data from Bangkok, we calculated the probability of two cases within a particular distance and found within a short time frame of each other being homotypic (i.e., caused by the same stereotype and therefore consistent with being transmission related) relative to the probability of any two cases being homotypic at that time. We were therefore able to characterize the spatial tendency for cases to be homotypic over various distances and interpret our findings as relative risk ratios. We found a strong spatial signal with case-pairs found within a month of each other being significantly more likely to be homotypic at distances up to 1km. There was a consistent pattern across each of the four serotypes. Further, when we incorporated temporal lags (i.e., looked at the impact of the presence of a case on future case distribution at that location), we found there existed significant spatial memory in case distribution. Clustering of homotypic case-pairs over short time periods of a few months was followed by inhibition over longer time-frames (i.e., fewer cases than expected) at similar spatial scales. By contrast there were fewer than expected heterotypic cases (cases caused by different serotypes) over time-frames of four to ten months. In addition, there was clustering of heterotypic cases at time lags of over 20 months. These patterns of homotypic and heterotypic spatial dependence are consistent with that expected from individual immunity patterns being reflected in the immunity of the local community, i.e., infection by a particular serotype causing (a) future protection from infection by homotypic viruses, (b) short-term protection from infection from heterotypic viruses and (c) increased risk of heterotypic infection over longer time frames [2–4].

The spatial extent at which cases of dengue tend to be found together will depend (to some extent at least) on the transmission kernel - i.e., how far sequential cases in a transmission chain tend to be from each other. We developed two approaches

to estimate mean transmission distances where we only have the space and time of where cases occur: a simpler form that uses the mean separation between observed cases at two time points that can be used in outbreak scenarios (chapter 3) and a more complex form for use in endemic settings where multiple transmission chains circulate (chapter 4). We demonstrated the robustness of these approaches through simulation and then applied them to case data. We found that despite our observation that cases of dengue tended to be found together at distances up to one km, the mean transmission distance was only 50m. This finding suggests that transmission between neighboring households is driving the spread of the pathogen in Bangkok. The major vector for dengue in Bangkok, *Aedes aegypti* doesn't travel very far and tends to reside within homes [5]. Human movements may also contribute to disease spread, especially during morning and evening periods when the mosquitoes are most active [6, 7]. We observed a similar transmission kernel for an outbreak of chikungunya, a disease also transmitted by the *Aedes* mosquito in Bangladesh, again supporting a major role for small scale transmissions between neighboring homes driving disease spread.

During an outbreak, we are usually most concerned with reducing disease risk across the population rather than identifying individuals directly at risk from a particular case. The home location of infections detected through passive surveillance can act as a marker of where infections are happening without needing to identify transmission related pairs. Our findings that there exists significant spatial dependence between dengue cases at distances up to one km, indicate that short transmission distances can create a much larger footprint of infection risk. Currently, mosquito control teams deployed by the local public health authority will spray 100m around the homes of dengue cases detected in Bangkok hospitals. While such a distance may capture directly transmitted infections, it represents only a small fraction of the total

number of infections occurring in that area. Our findings will inform mathematical models that look to optimize such intervention efforts.

Finally, it is important to note that our approaches capture global attributes of the disease systems we are investigating: the typical distance between directly transmission-related pairs of cases and the tendency for cases to be found together, irrespective of their transmission relationship. Implicit in our approaches is an assumption of stationarity, so for example the transmission kernel in one part of our study area is approximately the same as another part of the study area. There may, however, been neighborhood effects that impact local disease dynamics. Potential sources of neighborhood effects include population density differences, heterogeneities in housing types, mosquito habitats and differences in immigration and emigration that could affect immunity profiles. Our findings represent global estimates that average across such differences. Future work will consist of incorporating local factors as covariates and exploring spatial dynamics in subsets of the population.

Dengue and chikungunya are responsible for millions of infections each year. The methodological approaches and results presented here help us understand how the viruses are moving around communities and will aid tailor future intervention measures. While applied to two vector-borne pathogens, the methods will be applicable across infectious disease systems.

References

[1] A. Nisalak, T. P. Endy, S. Nimmannitya, S. Kalayanarooj, U. Thisayakorn, R. M. Scott, D. S. Burke, C. H. Hoke, B. L. Innis, and D. W. Vaughn, "Serotype-specific dengue virus circulation and dengue disease in Bangkok, Thailand from 1973 to 1999." *The American journal of tropical medicine and hygiene*, vol. 68, no. 2, pp. 191–202, Feb. 2003.

[2] S. B. Halstead, "Dengue," Imperial College Press, London, Oct. 2008.

[3] A. B. Sabin, "Research on dengue during World War II." *The American journal of tropical medicine and hygiene*, vol. 1, no. 1, pp. 30–50, Jan. 1952.

[4] D. S. Burke, A. Nisalak, D. E. Johnson, and R. M. Scott, "A prospective study of dengue infections in Bangkok." *The American journal of tropical medicine and hygiene*, vol. 38, no. 1, pp. 172–180, Jan. 1988.

[5] L. C. Harrington, T. W. Scott, K. Lerdthusnee, R. C. Coleman, A. Costero, G. G. Clark, J. J. Jones, S. Kitthawee, P. Kittayapong, R. Sithiprasasna, and J. D. Edman, "Dispersal of the dengue vector Aedes aegypti within and between rural communities." *The American journal of tropical medicine and hygiene*, vol. 72, no. 2, pp. 209–220, Feb. 2005.

[6] M. Yasuno and R. J. Tonn, "A study of biting habits of Aedes aegypti in Bangkok, Thailand." *Bulletin of the World Health Organization*, vol. 43, no. 2, pp. 319–325, 1970.

[7] S. T. Stoddard, B. M. Forshey, A. C. Morrison, V. A. Paz-Soldan, G. M. Vazquez-Prokopec, H. Astete, R. C. Reiner, S. Vilcarromero, J. P. Elder, E. S. Halsey, T. J. Kochel, U. Kitron, and T. W. Scott, "House-to-house human movement drives dengue virus transmission," *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

Part III

Appendices

Appendix A

Supplementary material to Chapter 2

A.1 Adapted space-time statistics

A.1.1 The space-time K-function

Spatiotemporal dependence is often characterized using the space-time K-function. The space-time K-function for multitypic point patterns describes spatiotemporal dependence as the expected number of points of type B, within a set distance (d) and time (t) from a point of type A multiplied by the expected space-time area occupied by each case of type B [1] [2] [3]:

$$K_{AB}(d,t) = \lambda_B^{-1} E[\#\ points\ of\ type\ B\ within\ d\ and\ t\ of\ a\ point\ of\ type\ A] \quad \text{(A.1)}$$

where λ_B is the spatiotemporal intensity of the points of type B.

The space-time K function is typically estimated as:

$$\hat{K}_{AB}(d,t) = \frac{SL \sum_{i=1} \sum_{j=1} I(j \in \Omega_i(d,t) | i \in A, j \in B)}{n_A n_B} \quad \text{(A.2)}$$

where n_A is the total number of points of type A and $\Omega_i(d,t)$ is the set of points within distance, d, and time, t, of i.

The space-time K-function can be use to define $D_{0AB}(d,t)$, the relative difference in the probability of observing a point of type B within d and t from a point of type A to that expected given the separate clustering observed in space and that observed in time [1] [2] [3] [4] [5].

$$D_{0AB}(d,t) = \frac{K_{AB}(d,t)}{K_{AB}(d,\cdot)K_{AB}(\cdot,t)} - 1 \tag{A.3}$$

$$= \frac{Pr(j \in \Omega_i(d,t) | i \in A, j \in B)}{Pr(j \in \Omega_i(d,\cdot) | i \in A, j \in B) Pr(j \in \Omega_i(\cdot,t) | i \in A, j \in B)} - 1 \tag{A.4}$$

A.1.2 Extending to space-time windows

The space-time K-function and D_0 are calculated cumulatively in respect of space and time. This results in crude characterization and may fail to detect changing patterns of clustering. If, for example, all disease transmission is occurring within the home, the space-time K-function may be unable to accurately characterize spatial dependence beyond the typical dimensions of a home (Figure A.1).

APPENDIX A.

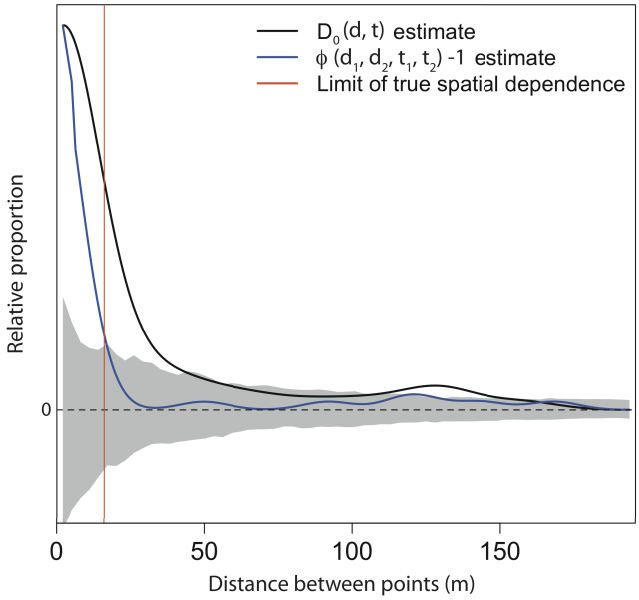


Figure A.1: $D_0(d,t)$ and $\phi(d_1,d_2,t_1,t_2)-1$ estimates from a basic simulation of household disease transmission. Index cases occur completely spatially at random at a randomly chosen time during a 10 day period. Each index case has one secondary case that occurs within 20m a mean of 2 days later. Both analyses use a temporal range of 2 days ($t = 2$ days for D_0 and $t_1 = 0$ and $t_2 = 2$ days for ϕ). The mean spatial window size for $\phi(d_1,d_2,t_1,t_2)$ is kept at 10m (so $d_2 - d_1 = 10$m) . The shaded area represents 95% limits from a null distribution for D_0 from 1000 simulations where the spatial location of all cases are randomly reassigned with the time label held fixed (as suggested by [3]).

To appropriately characterize the spatiotemporal scale of dependence and allow for changing patterns of dependence, we introduce a space-time window within which spatiotemporal dependence is calculated.

$$K_{AB}(d_1,d_2,t_1,t_2) = \lambda_B^{-1} E[\#\ points\ of\ type\ B\ within\ d_1\ and\ d_2 \\ and\ t_1\ and\ t_2\ of a\ point\ of\ type\ A] \qquad (A.5)$$

A.1.3 Extending to relationships between points

$K_{AB}(d,t)$ reflects the spatiotemporal dependence between two sets of points, those of type A and those of type B. It may also be of interest to understand the dependence between sets of points that are related through a function. An example is points that have the same label (such as serotype for homotypic dependence) or points that have different labels (heterotypic dependence). We create a general form of the space-time K-function:

$$K_{f()}(d_1,d_2,t_1,t_2) = \lambda_j^{-1} E[\#\ points\ j\ within\ d_1\ and\ d_2 \\ and\ t_1\ and\ t_2\ of\ any\ point\ i | f(i,j)\ is\ true] \qquad (A.6)$$

A.2 Characterizing short term spatial dependence

A.2.1 the τ function

To describe spatial dependence of homotypic cases occurring within the same month we calculate the probability that cases occur within the same month and within a defined spatial window are homotypic relative to the probability that any two cases are homotypic within that month, irrespective of spatial location:

$$\tau(d_1,d_2) = \frac{Pr(z_{ij}=1 | j \in \Omega_i(d_1,d_2))}{Pr(z_{ij}=1 | j \in \Omega_i(\cdot))} \qquad (A.7)$$

where z_i is the serotype of case i and $\Omega_i(d_1,d_2)$ is the set of cases that occur within the same month and within distances d_1 and d_2 of case i.

As both the numerator and the denominator are calculated with respect to cases

that occur within the same month, temporal clustering of cases over the timeframe of the dataset, such as due to seasonal forcing, do not affect the estimates. In addition, as we are calculating the probability that any two cases are homotypic within a set distance range, any spatial clustering that occurs to all cases irrespective of serotype will also not affect our estimates. In such a way underlying factors that could create spatial clustering themselves, including hospital utilization rates, population density differences and distances of homes from health facilities, which affects all cases, do not change our estimates.

We estimate τ as:

$$\hat{\tau}(d_1, d_2) = \frac{\sum_{i=1}^{N} \sum_{j \in \Omega_i(d_1,d_2)} z_{ij}}{\sum_{i=1}^{N} |\Omega_i(d_1, d_2)|} / \frac{\sum_{i=1}^{N} \sum_{j \in \Omega_i(\cdot)} z_{ij}}{\sum_{i=1}^{N} |\Omega_i(\cdot)|} \tag{A.8}$$

$\Omega_i(\cdot)$ is the set of all points occurring within all distances within the same month from point i.

A.2.2 τ for individual serotypes

We calculate the probability that a case caused by a specified serotype, z, is found within a defined spatial window and within the same month as another case caused by z, relative to the probability of them being found within the same month, irrespective of spatial location.

$$\tau_z(d_1, d_2) = \frac{Pr(z_{ij} = 1 | j \in \Omega_i(d_1, d_2), z_i = z)}{Pr(z_{ij} = 1 | j \in \Omega_i(\cdot), z_i = z)} \tag{A.9}$$

A.2.3 Null distribution calculation for τ

We use a Monte Carlo procedure to create a null distribution to test for the significance of the label (serotype) attached to each case. The null hypothesis is that the serotype of the case is independent to the spatial location of the cases within any month. Ninety-five percent null distributions are constructed by randomly reassigning the serotype label over 1000 iterations for each d_1, d_2 window, keeping the total number for each label the same within any month.

A.2.4 Confidence intervals for τ

Confidence intervals are calculated by bootstrapping. The sampling unit for the bootstrap is the individual point locations. Ninety-five percent confidence intervals are calculated from the 2.5% and 97.5% percentiles of 1000 bootstrap samples of $\hat{\tau}(d_1, d_2)$ for each d_1, d_2 window.

A.2.5 Temporal extension of the τ function

The temporal extension of our τ function estimates the probability of two cases that occur within t_1, t_2 and d_1, d_2 of each other being homotypic relative to the probability of any two cases being homotypic within that temporal window ($\tau(d_1, d_2, t_1, t_2)$). As the numerator and the denominator are calculated only within the t_1, t_2 window, secular changes across the entire study period do not affect the estimates.

$$\tau(d_1, d_2, t_1, t_2) = \frac{Pr(z_i = z_j | i \in \Omega_j(d_1, d_2, t_1, t_2))}{Pr(z_i = z_j | i \in \Omega_j(\cdot, t_1, t_2))} \tag{A.10}$$

A.2.6 The space-time K-function

$\tau(d_1, d_2, t_1, t_2)$ is equivalent to the ratio of the windowed space-time K-functions (see equation (3.5)]) for homotypic cases and that for all cases, irrespective of serotype (denoted here as $K_{(\cdot)}$), divided by the equivalent ratio of the same K-functions across all space.

$$\tau(d_1, d_2, t_1, t_2) = \frac{Pr(i \in \Omega_j(d_1, d_2, t_1, t_2) | z_i = z_j)}{Pr(i \in \Omega_j(d_1, d_2, t_1, t_2))} \frac{Pr(i \in \Omega_j(\cdot, t_1, t_2))}{Pr(i \in \Omega_j(\cdot, t_1, t_2) | z_i = z_j)} \quad \text{(A.11)}$$

$$= \frac{K_{f()}(d_1, d_2, t_1, t_2)}{K_{(\cdot)}(d_1, d_2, t_1, t_2)} / \frac{K_{f()}(\cdot, t_1, t_2)}{K_{(\cdot)}(\cdot, t_1, t_2)} \quad \text{(A.12)}$$

where f() describes the points that are the same serotype as each other.

A.3 Characterizing longer term spatiotemporal dependence

A.3.1 The general ϕ function

To describe spatiotemporal dependence over longer time periods we calculate the probability that point j, whose serotype bears relation $f()$ to point i is within a set spatiotemporal window from i, relative to that expected if the spatial and the temporal dependence between them were independent.

$$\phi_{f()}(d_1, d_2, t_1, t_2) = \frac{Pr(j \in \Omega_i(d_1, d_2, t_1, t_2) | f(i, j))}{Pr(j \in \Omega_i(d_1, d_2, \cdot) | f(i, j)) Pr(j \in \Omega_i(\cdot, t_1, t_2) | f(i, j))} \quad \text{(A.13)}$$

The $\phi_{f()}(d_1, d_2, t_1, t_2)$ is related to $D_{0f()}(d_1, d_2, t_1, t_2)$ such that:

$$\phi_{f()}(d_1, d_2, t_1, t_2) = D_{0f()}(d_1, d_2, t_1, t_2) + 1 \tag{A.14}$$

A.3.2 The ϕ function for homotypic and heterotypic spatiotemporal dependence

In the case of characterizing homotypic spatiotemporal dependence ($\phi_{hom}(d_1, d_2, t_1, t_2)$), $f(i, j) = hom(i, j)$ and is true when i and j are of the same serotype. In the case of heterotypic spatiotemporal dependence ($\phi_{het}(d_1, d_2, t_1, t_2)$), $f(i, j) = het(i, j)$ and is true when i and j are of different serotypes.

We estimate $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\phi_{het}(d_1, d_2, t_1, t_2)$ as:

$$\hat{\phi}_{hom}(d_1, d_2, t_1, t_2) = \frac{(\sum_{i=1}^{N} \sum_{j \in \Omega_i(d_1, d_2, t_1, t_2)} z_{ij})(\sum_{i=1}^{N} \sum_{j \in \Omega_i(\cdot, \cdot)} z_{ij})}{(\sum_{i=1}^{N} \sum_{j \in \Omega_i(d_1, d_2, \cdot)} z_{ij})(\sum_{i=1}^{N} \sum_{j \in \Omega_i(\cdot, t_1, t_2)} z_{ij})} \tag{A.15}$$

$$\hat{\phi}_{het}(d_1, d_2, t_1, t_2) = \frac{(\sum_{i=1}^{N} \sum_{j \in \Omega_i(d_1, d_2, t_1, t_2)} (1 - z_{ij}))(\sum_{i=1}^{N} \sum_{j \in \Omega_i(\cdot, \cdot)} (1 - z_{ij}))}{(\sum_{i=1}^{N} \sum_{j \in \Omega_i(d_1, d_2, \cdot)} (1 - z_{ij}))(\sum_{i=1}^{N} \sum_{j \in \Omega_i(\cdot, t_1, t_2)} (1 - z_{ij}))} \tag{A.16}$$

where z_{ij} is equal to 1 if the serotypes are the same and 0 otherwise.

Values of $\phi_{hom}(d_1, d_2, t_1, t_2)$ above 1 indicate a greater tendency for cases caused by the same serotype to be found together within the spatial and temporal ranges than would be expected if the clustering observed in space was independent to the clustering observed in time. As well as allowing the detection of changing patterns of spatiotemporal dependence, this approach allows for separate dynamics at the same temporal range in the homotypic and heterotypic analyses.

A.3.3 Note on underlying spatial and temporal clustering

$\phi_{f()}(d_1, d_2, t_1, t_2)$ calculates the probability of two related points being found near each other in both space and time relative to that expected if overall spatial and overall temporal dependence were independent. Any clustering processes that occur in space or time, such as from seasonal factors or population density differences, affect both the numerator and the denominator of $\phi_{f()}(d_1, d_2, t_1, t_2)$ and therefore do not affect our estimates.

A.3.4 Confidence intervals for ϕ

Ninety-five percent confidence intervals are generated through bootstrapping. The sampling unit for the bootstrap is the individual point locations. We perform 1000 bootstrap samples for each combination of d_1, d_2, t_1, t_2. Confidence plots are generated by calculating the proportion of the bootstrapped simulations that are greater than 1.0. A proportion greater than 97.5% indicates significant positive spatiotemporal dependence between cases whereas a proportion less than 2.5% indicates significant negative spatiotemporal dependence.

A.4 Simulations to illustrate robustness of $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$

In order to explore the robustness of $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ to variability in reporting patterns and to different underlying population structures, we simulated disease transmission processes.

A.4.1 Model structure

We constructed an individual based spatially explicit transmission model of all four serotypes of dengue. The five different model constructions are set out in Table 1. Individuals were either randomly distributed or inhomogeneously distributed (Figure 2).

Model	Population density distribution	Seasonal difference in incidence	Spatial dependence between infector and infected
A	Homogeneous	No	No
B	Inhomogeneous	No	No
C	Homogeneous	Yes	No
D	Inhomogeneous	Yes	No
E	Homogeneous	No	Yes

Table A.1: Different population and seasonality assumptions for the five simulation models.

The infectious process is modeled by initially infecting 100 randomly chosen individuals (25 each with each of the 4 serotypes). Transmission events are modeled by randomly identifying individuals in a spatially dependent manner for some simulations and without spatial dependence in another (as described below). The time of transmissions are normally distributed with mean of 14 days (representing the time between successive human infections). We assume the system is at equilibrium and therefore has an effective reproductive number of one. The location, time and serotype of infection at each transmission event is recorded.

A.4.2 Effect of population structure and seasonality

We investigate whether underlying differences in the (a) population density or (b) differences in the seasonality of infections could induce the appearance of spatiotemporal dependence between cases in $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ where no true dependence exists.

We simulate a disease simulation process where infections occurred in randomly chosen individuals with no spatial dependence between the infector and the infectee. We run the simulations in homogeneous (models A and C) and inhomogeneous population density structures (models B and D) (Figure A.2). In both population density structure scenarios we also vary the seasonality of cases with either no seasonality in the effective reproductive number (models A and B) or with seasonal variation in transmission present (models C and D), where we use:

$$R_t = 1 + 0.05sin\left(\frac{2\pi t}{365}\right) \tag{A.17}$$

where R_t is the effective reproductive number at time t.

We calculate $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ for each simulation.

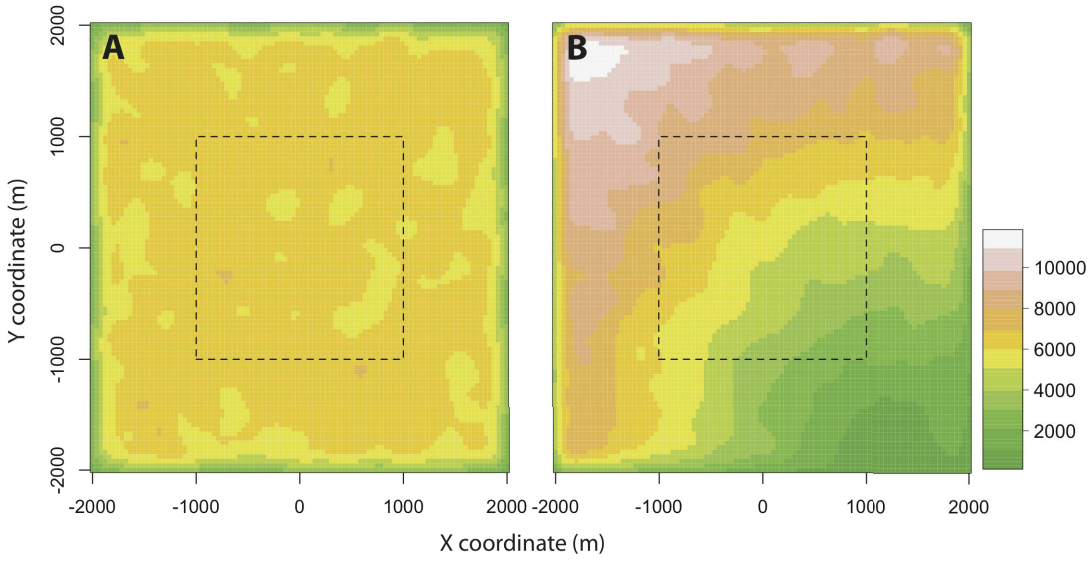


Figure A.2: Population density distribution for the simulations. Population density distribution in number of individuals per km2 of simulation using (A) homogeneous and (B) inhomogeneous population structures. The dotted line represents the region in which the clustering statistics were calculated. This avoids the effects of infections outside the simulation area.

We found that even when there existed strong seasonality or spatial dependence in the structure of the underlying population (but not the infection process), both the $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ statistics correctly identified no spatial dependence between cases (Figure A.3).

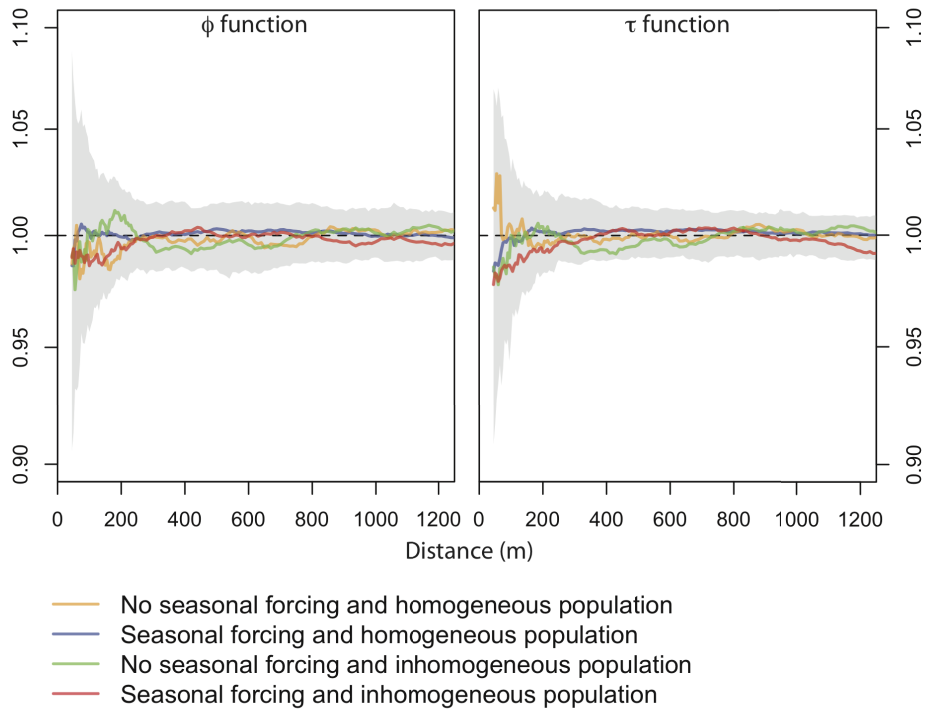


Figure A.3: Effect of population structure and seasonal forcing.$\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ estimates under different population structures and seasonal forcing assumptions where there exists no true spatiotemporal dependence in the simulated transmission process. For $\phi_{hom}(d_1, d_2, t_1, t_2)$, $t_1 = 0$ and $t_2 = 1$ month. The grey shaded region represents 95% intervals for the null distribution where there exists seasonal forcing in an inhomogeneous population.

A.4.3 Effect of reporting bias

To assess whether reporting biases could effect the estimated spatiotemporal clustering from $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ we simulate a disease transmission process where sequential cases in a transmission chain has a mean spatial separation of 100 m (model E). We compare the results in $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ when all cases are used in the calculations to the following scenarios where only a subset of cases

are observed:

(a) Completely spatially random reporting (all cases, irrespective of location, has an equal probability of being reported).

(b) Spatially-dependent reporting (the probability of being reported depends on spatial location). E.g. the probability of turning up at a specific hospital may depend on the distance to the hospital. The probability of report is modeled using:

$$P(\text{case reported}) = exp(-\bar{d}) \tag{A.18}$$

where $\bar{d}$ is the distance from the centre of the polygon, normalized by the maximum distance of a case from the hospital.

(c) Temporally-dependent reporting (the probability of being reported depends on when the case occurred). This may occur in situations where hospitals are more likely to detect dengue during certain times of the year. The probability of report is modeled using:

$$P(casereported) = 0.5 + 0.5sin\left(\frac{2\pi t}{365}\right) \tag{A.19}$$

(d) Spatially- and temporally- dependent reporting (the probability of being reported depends on where and when the case occurs but the processes that determines the spatial and temporal thinning are independent of each other). The probability of a case being reported is calculated by multiplying the probabilities in (b) and (c) above.

(e) Spatiotemporal-dependent reporting (as (d) but the processes that determines the spatial and temporal thinning are linked). This may occur if there are health infrastructure changes during the period of the observations that alter the probability that a dengue case attends a particular hospital during the period of analysis. The probability of a case being reported is 0.1 for all cases that occur in the second half of the time series that are over the median distance away from the centre of the area and 0.9 for all other cases.

We found that biased reporting that did not depend on when or where the case occurs did not bias the estimates of $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ (Figure A.4).

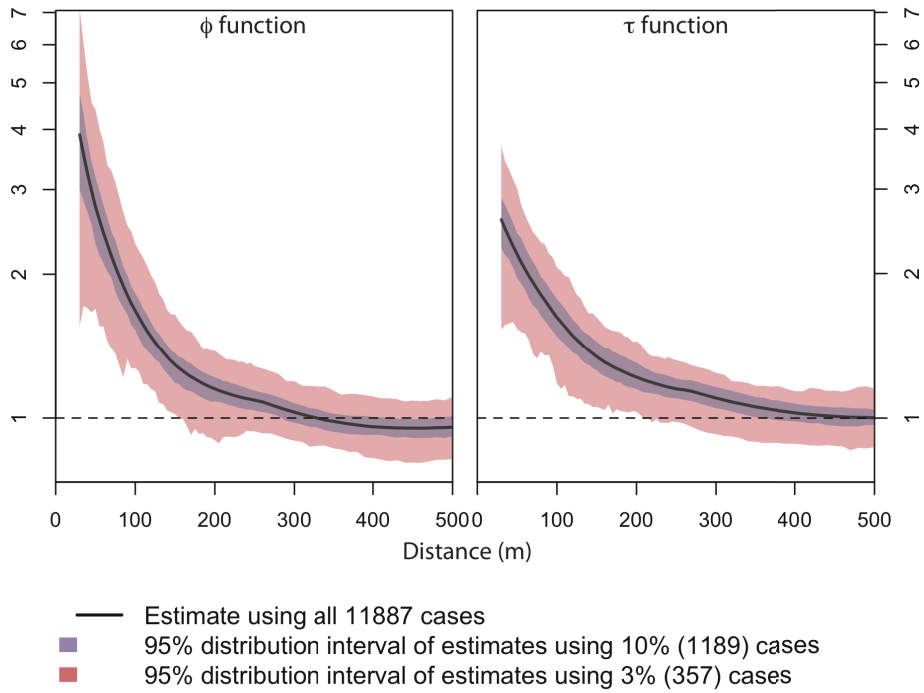


Figure A.4: $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ estimates under spatially random reporting. For $\phi_{hom}(d_1, d_2, t_1, t_2)$, $t_1 = 0$ and $t_2 = 1$ month. The shaded regions represent 95% distributions of the estimates when a random subset of all cases is observed.

In addition where there exists, spatially-dependent reporting, temporally-dependent

reporting and spatially- and temporally- dependent reporting, there is also no bias in $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ (Figure A.5). In the event of spatiotemporal-dependent reporting, where the spatial and temporal processes that determine if a case is reported are linked, $\phi_{hom}(d_1, d_2, t_1, t_2)$ is biased but only to a small extent. In this scenario $\tau(d_1, d_2)$ is not biased.

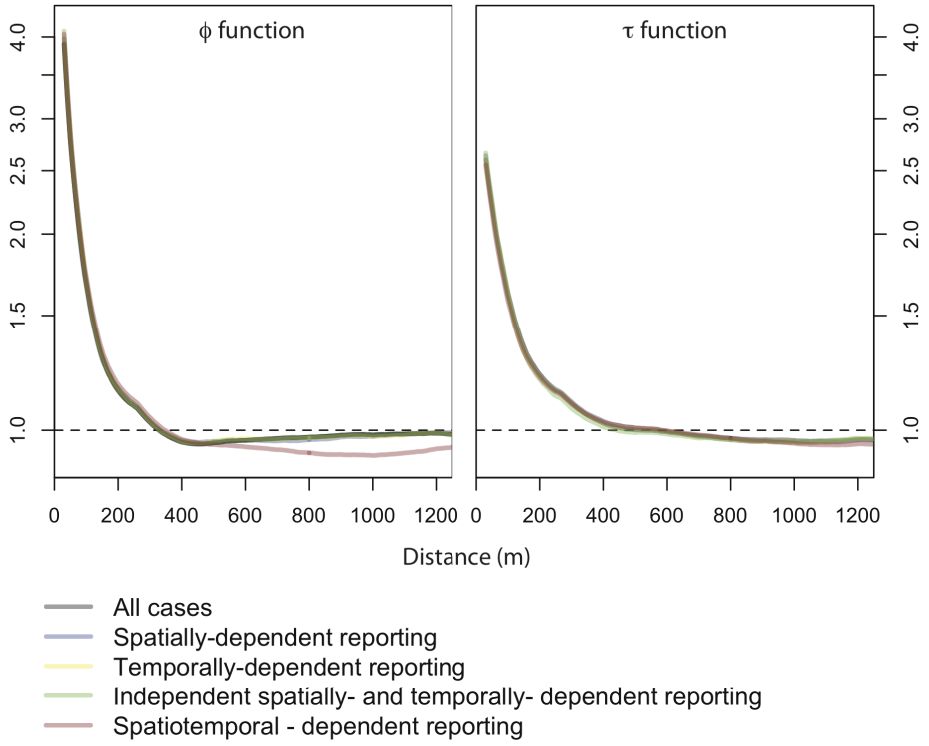


Figure A.5: $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ estimates under different types of reporting bias. A disease transmission process was simulated and a random binomial distribution was then used to determine which cases from the parent population were reported. The probability of a particular case being reported depended on its location and when it occurred. This probability differed by the type of reporting bias. The lines represent mean $\phi_{hom}(d_1, d_2, t_1, t_2)$ and $\tau(d_1, d_2)$ from 500 repeats of the reporting procedure, using the same parent population of cases each time.

A.5 Sensitivity analyses

We conduct various sensitivity analyses to explore the consistency of our results to differences in spatial location, differences in the spatial window of analysis and aggregation of data.

A.5.1 Geographical differences in the short-term spatial clustering

To understand if there are differences in the clustering observed in different parts of the city, we calculate individual $\tau(d_1, d_2)$ estimates for cases coming from the north, south, east and west of the hospital separately. We find consistent patterns in spatial dependence across the different locations (Figure A.6).

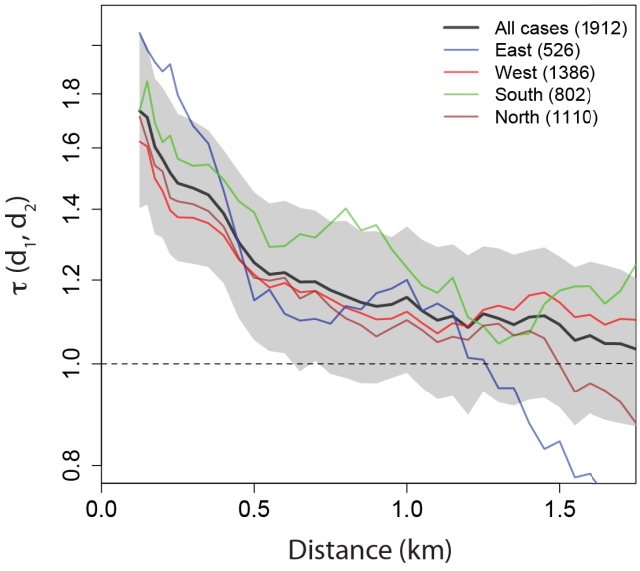


Figure A.6: Homotypic spatial dependence estimates for different areas of the city relative to the hospital. The number of cases that are located in that region and used in the analysis are in parentheses. The grey shaded region is 95% bootstrap confidence interval using all cases.

A.5.2 Window of analysis

The main anaysis uses 500m as the spatial window of analysis (i.e. the difference between d_2 and d_1 in $\tau(d_1, d_2)$). To explore the sensitivity of our results to different sized windows we repeat the analyses, varying the window size from 250m to 3km.

We found that increasing the size of the window reduces the variability in the $\tau(d_1, d_2)$ estimates (Figure A.7). This increase in smoothness of the estimates is due to more data points being included in each single estimate as the spatial area of analysis is increased. However, wider windows also reduce the ability to detect small changes in spatial dependence (Figure A.1).

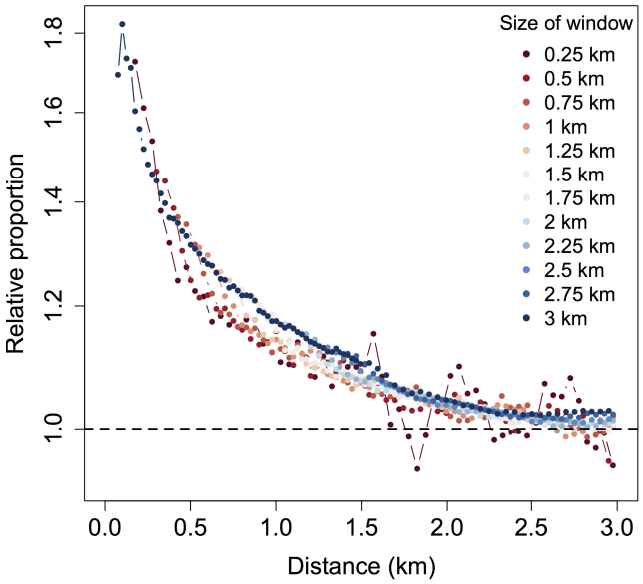


Figure A.7: Impact of window size on estimates of $\tau(d_1, d_2)$. Window size (difference between d_2 and d_1) varied between 0.25km and 3km.

A.5.3 Aggregation of data

It is often difficult to collect exact spatial data. Addresses may only be available at coarser resolutions such as city block or zip code. To assess the consistency of our results to different levels of spatial aggregation we repeat the estimation of the short-term clustering with addresses aggregated to different spatial scales. To assign the new (aggregated) spatial location for each point, we place a fine spatial grid over the city, where the distance between each grid cell is either 100m, 500m, 1km or 5km and identify the closest grid cell for each case. We then re-estimate $\tau(d_1, d_2)$ for each level of spatial aggregation.

We found that aggregating points by even up to 1km produced broadly consistent

results (Figure A.8).

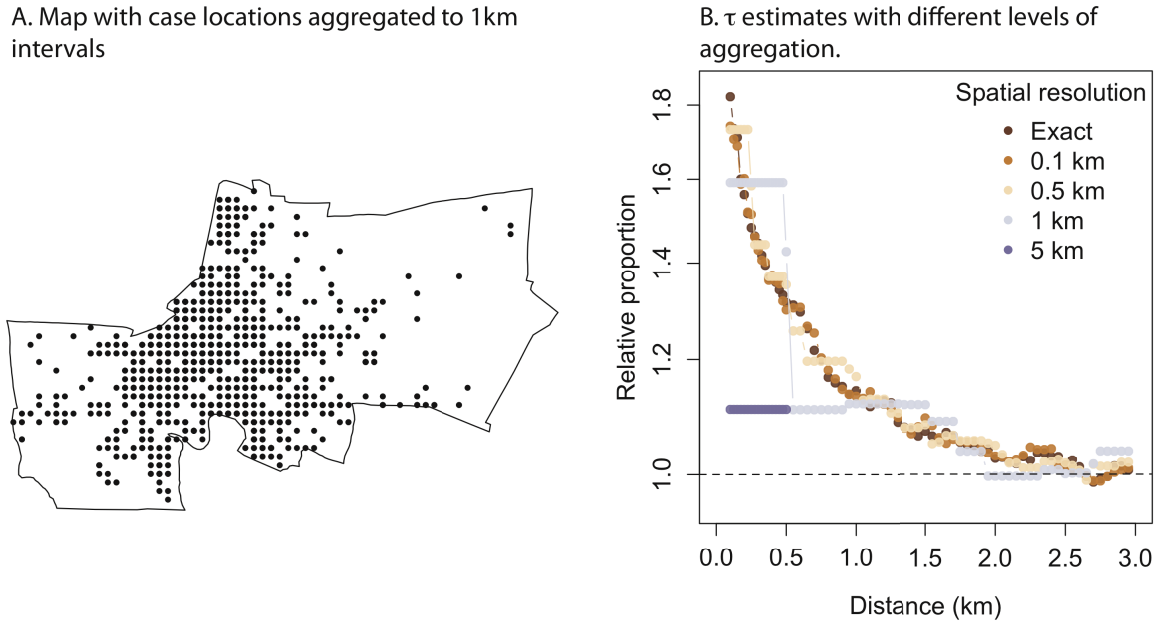


Figure A.8: Impact of spatial aggregation of data on estimates of $\tau(d_1, d_2)$. Addresses were reassigned to grid cells in the city, where the distance between cells was varied between 0.1km and 5km. The map on the left shows the aggregation of addresses for a resolution of 1km. The plot on the right shows the estimates of $\tau(d_1, d_2)$ for the different levels of aggregation.

A.6 Data analysis software

Data analysis was performed using R 2.12.1.

References

[1] B. D. Ripley, "The Second-Order Analysis of Stationary Point Processes," *Journal of Applied Probability*, vol. 13, no. 2, pp. 255–266, Jun. 1976.

[2] A. C. Gatrell, T. C. Bailey, P. J. Diggle, and B. S. Rowlingson, "Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology," *Transactions of the Institute of British Geographers, New Series*, vol. 21, no. 1, pp. 256–274, Jan. 1996.

[3] P. J. Diggle, A. G. Chetwynd, R. Häggkvist, and S. E. Morris, "Second-order analysis of space-time clustering." *Statistical methods in medical research*, vol. 4, no. 2, pp. 124–136, Jun. 1995.

[4] N. P. French, H. E. McCarthy, P. J. Diggle, and C. J. Proudman, "Clustering of equine grass sickness cases in the United Kingdom: a study considering the effect of position-dependent reporting on the space-time K-function." *Epidemiology and infection*, vol. 133, no. 2, pp. 343–348, Apr. 2005.

[5] H. J. Lynch and P. R. Moorcroft, "A spatiotemporal Ripley's K-function to analyze interactions between spruce budworm and fire in British Columbia, Canada," *Canadian Journal of Forest Research*, vol. 38, no. 12, pp. 3112–3119, 2008.

Appendix B

Supplementary material to Chapter 4

	Positive	Negative	Total tested	Not tested	Total
Symptoms	172	74	246	201	447
No Symptoms	52	119	172	1351	1523
Total	225	193	418	1544	1970

Table B.1: Number of individuals interviewed and tested from the three villages in the outbreak investigation in 2012.

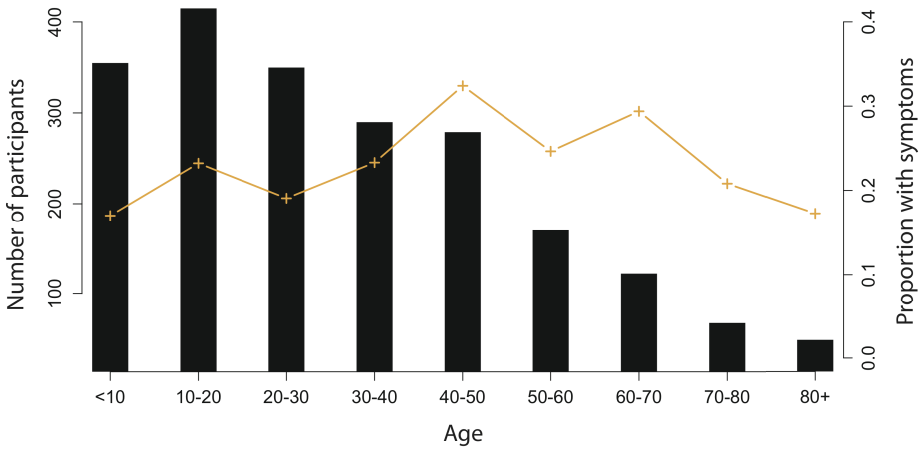


Figure B.1: Number of participants by age and the proportion that reported symptoms consistent with chikungunya.

Appendix C

Supplementary material to Chapter 5

C.1 Detailed methods

C.1.1 PRNT calculation

The plaque reduction proportion for each dilution (pd) was calculated as:

$$p_d = 1 - \frac{n_d}{n_0} \tag{C.1}$$

where n_d number of plaques at dilution d and n_0 is the number of plaques formed when no sera is added. PRNT_{50} can be estimated using generalized linear regression using the log-transformed dilutions and either a probit, logit or complementary log-log (cloglog) link function:

APPENDIX C.

(1) probit:

$$\text{probit}(p_d) = \phi^{-1}(p_d) \tag{C.2}$$

(2) logit:

$$\text{logit}(p_d) = log\left(\frac{p_d}{1-p_d}\right) \tag{C.3}$$

(3) complementary loglog:

$$\text{cloglog}(p_d) = log(-log(1-p_d)) \tag{C.4}$$

where ϕ^{-1} is the inverse cumulative distribution function of the standard normal distribution. These regressions are used to interpolate the titer at which defined reductions (PRNT cut-points) occur from the observed reductions (e.g. a 50% reduction for a cut-point of $PRNT_{50}$). Variability in plaque counts may result in the number of plaques counted under high dilutions exceeding the number formed when no sera is added. To avoid errors in the transformations, values of pd less than 0 were replaced with 0.001. As some laboratories use different cut-points, we also calculated PRNTs over the range $PRNT_{40}$ to $PRNT_{90}$.

As an alternative approach, some laboratories use non-linear regression approaches. A popular method is the four-parameter model used by Prism 6 software (La Jolla, CA) for sigmoidal curves, which finds optimum values for the maximum and minimum plaque reductions, the slope of the linear part of the curve and the dilution of the inflexion point. To use this approach researchers need the plaque reductions from

at least four dilutions.

C.1.2 Bias and Mean Squared Error calculation

For each experiment, we calculated PRNT using each of the models (probit, logit, cloglog regression and non-linear regression). Bias was suggested when there was a systematic difference between the PRNT estimates using the model and PRNTsp:

$$bias^{i}(x, v, p, m) = PRNT^{i}(x, v, p, m) - PRNT^{sp}(x, v, p) \tag{C.5}$$

where $PRNT^{i}(x, v, p, m)$ is the PRNT estimate from experiment i conducted with viral strain v (the five virus strains in Table 5.1) in serum pool p (either the high titer or the lower titer pool) estimated using model m (probit, logit, cloglog or four-parameter non-linear regression models) at a PRNT cut-point of x (varied from 40 to 90); $PRNT^{sp}(x, v, p)$ is the estimate of the truetiter, where x is the PRNT cut-point, v is the viral strain and p is the serum pool.

In addition, for each viral strain, for both the high and low titer pools, we calculated the mean squared error (MSE) of each parametric model using the following relationship:

$$MSE(x, v, p, m) = \text{Variance}(x, v, p, m) + \text{Bias}(x, v, p, m)^2 \tag{C.6}$$

Where Variance(x, v, p, m) is the variance and Bias is the mean bias in PRNT estimates using model m from all experiments conducted with viral strain v in serum

pool p. We reported an average MSE, bias ($\bar{B}(x,m)$) and variance ($\bar{V}(x,m)$) for each cut-point and model, weighted by the number of experiments using each virus and serum pool.

C.1.3 Confidence interval calculation

We used the bias and variance estimates to calculate 95% asymptotic confidence intervals for an example true titer of 1:300 using a cut-point of x and model m:

$$10^{log_{10}(300)+\bar{B}(x,m)\pm 1.96\sqrt{(V)(x,m)}} \tag{C.7}$$

The confidence interval can be interpreted as the range of values that contain 95% of measured titers when the true titer is 1:300.

C.1.4 Multilevel model

We constructed a multilevel model with a random intercept for each viral strain and serum pool combination (listed in Table 1):

$$y_{ij} = \gamma_0 + \mu_j + \beta_1 \text{Pass}_i + \beta_2 \text{Cell}_i + \beta_3 \text{Age}_i \tag{C.8}$$

where y_{ij} is the log-transformed PRNT_{50} estimate (using probit regression) for experiment i using serum pool j, μ_j represents the random intercept for serum pool j, Pass is the total number of passages, Cell is a factor that represents either passages

in C6/36 and LLC-MK2 cells, C6/36 and SM cells or C6/36 cells only. Age is the age of the virus stock at the time of the experiment (in years). We assumed the errors in the model were normally distributed.

C.2 Bias by experiments using probit model

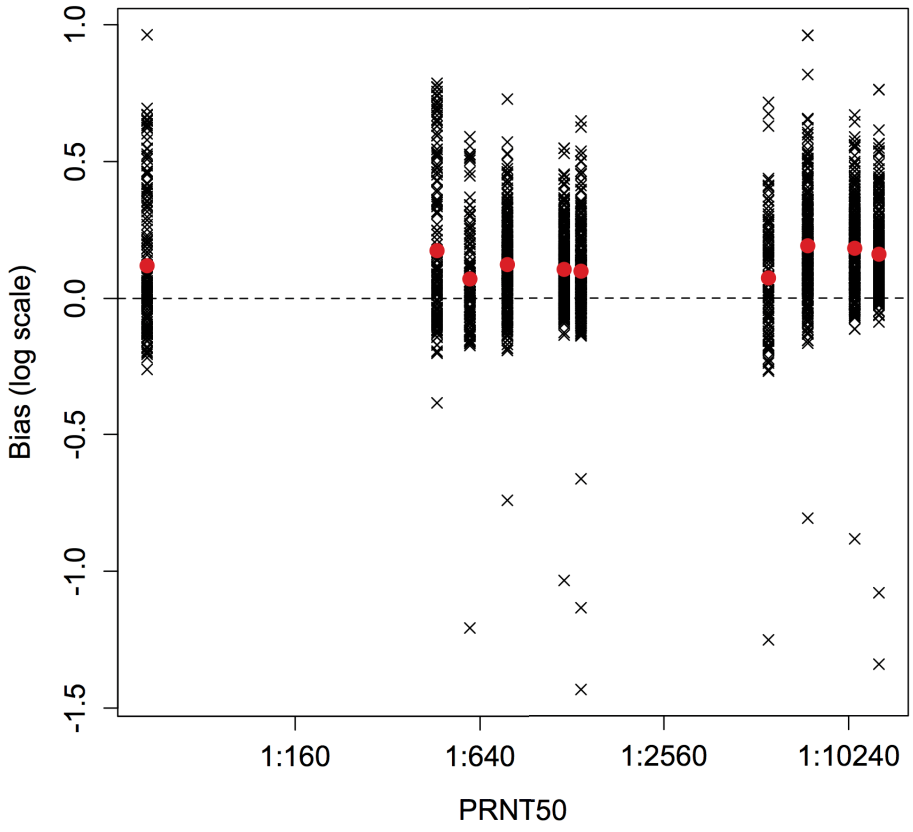


Figure C.1: Bias in $PRNT_{50}$ from all experiments using a conventional probit transformation by titer (log10 scale). The red dots represent the mean bias from each serum pool.

Curriculum Vitae

Henrik Salje
E6035, Johns Hopkins School of Public Health
615 N. Wolfe Street
Baltimore MD, 21205
hsalje@jhsph.edu

Education

2009 - 2014	**Johns Hopkins School of Public Health, Baltimore**
	Ph. D. Epidemiology
	M.H.S. Biostatistics
1998 - 2002	**Oxford University**
	BSc/MSc Biochemistry

Professional Experience

Research

2009 - 2014	**Johns Hopkins School of Public Health, Baltimore**
	Department of Epidemiology
	Graduate Research Assistant

Non-Research

2008 - 2009	**The Delphi Partnership**
	Co-founder and Finance Director
2004 - 2008	**Dawnay, Day Investment Banking**
	Associate Director
2003 - 2004	**Ernst & Young Corporate Finance**
	Analyst

Teaching Experience

Johns Hopkins Bloomberg School of Public Health, Baltimore

2013	Teaching Assistant - infectious disease dynamics
2010 - 2013	Teaching Assistant - GIS and spatial statistics
2010 - 2013	Teaching Assistant - advanced spatial statistics

International Centre for Diarrhoeal Disease Research, Bangladesh

2013	Instructor - spatial statistics

Publications

H. Salje et al., Revealing the microscale spatial signature of dengue transmission and immunity in an urban population, *PNAS*, 2012.

H. Salje et al., Impact of neighborhood biomass cooking patterns on episodic high indoor particulate matter concentrations in clean fuel homes in Dhaka, Bangladesh. *Indoor Air*, 2013.

A. Sun, M. Pai, H. Salje et al., Modeling the Impact of Alternative Strategies for Rapid Molecular Diagnosis of Tuberculosis in Southeast Asia. *American Journal of Epidemiology*, 2013.

E. Gurley, H. Salje et al., Indoor exposure to particulate matter and age at first acute lower respiratory infection in a low-income, urban community in Bangladesh. *American Journal of Epidemiology*, 2014 (in press)

E. Gurley, N. Homaira, H. Salje et al. Indoor exposure to particulate matter and the incidence of acute lower respiratory infections among children: a birth cohort study in urban Bangladesh. *Indoor Air*, 2013.

E. Gurley, H. Salje et al., Seasonal concentrations and determinants of indoor particulate matter in a low-income community in Dhaka, Bangladesh, *Environmental Research*, 2012.

M. Lomax, H. Salje et al., 8-OxoA Inhibits the Incision of an AP Site by the DNA Glycosylases Fpg, Nth and the AP Endonuclease HAP1, *Radiation Research*, 2005.

Under review

H. Salje et al., Variability in dengue titer estimates from plaque reduction neutralization tests poses a challenge to epidemiological studies and vaccine development

H. Salje et al., Does deployment strategy matter? Modeling the scale-up of novel tuberculosis diagnostics in the Indian healthcare system

S. Khan, H. Salje et al., Characterization of Japanese encephalitis dynamics among domestic pigs in northwest Bangladesh and the potential impact of pig vaccination

P. Bhoomiboonchoo, R. Gibbons, A. Huang, I Yoon, D. Buddhari, A. Nisalak, N. Chansatiporn, M. Thipayamongkolgul, S. Kalanarooj, T. Endy, A. Rothman, A. Srikiatkhachorn, S. Green, M. Mammen, D. Cummings & H. Salje, Using gravity models to describe dengue dynamics in Kamphaeng Phet, Thailand.

Professional Activities

- Reviewer, *American Journal of Epidemiology* 2013
- Reviewer, *PLoS Neglected Tropical Diseases* 2013 - 2014
- Reviewer, *BMC Infectious Diseases* 2013 - 2014
- Reviewer, *Journal of Health, Population and Nutrition* 2012

Funding and Awards

- NSF Doctoral Dissertation Improvement Grant (Co-PI, 2012 - 2014). Using Phylogeography To Understand the Spatiotemporal Clustering of Dengue Cases in Bangkok.
- Lilienfeld Prize, Society of Epidemiological Research (2012). Best paper in epidemiology.
- Louis I. and Thomas D. Dublin Award (2012). For the advancement of epidemiology and biostatistics.
- JHSPH Department of Epidemiology Research Grant (2012). Characterization of dengue transmission dynamics in Bangladesh.
- Global Health Field Research Award (2011). Phylogeography of dengue in Thailand.
- Sommer Scholarship (2009 - 2014).